AD_____

Award Number: DAMD17-97-1-7058

TITLE: Development of an Integrated Program of Health Related Quality of Life Research for the National Surgical Adjuvant Breast and Bowel Project (NSABP)

PRINCIPAL INVESTIGATOR: Richard D. Day, Ph.D.

CONTRACTING ORGANIZATION: University of Pittsburgh
Pittsburgh, Pennsylvania 15260

REPORT DATE: September 2002

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 2002	Final (1 September 1997 - 31 August 2002)

4. TITLE AND SUBTITLE
Development of an Integrated Program of Health-Related Quality-of-Life Research for the National Surgical Adjuvant Breast and Bowel Project (NSABP)

5. FUNDING NUMBER
DAMD17-97-1-7058

6. AUTHOR(S)
Richard D. Day, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Pittsburgh
Pittsburgh, Pennsylvania 15260
email DAY@NSABP.PITT.EDU

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

This Career Development Award (CDA) was specifically intended to support Dr. Day in the development of a Health-Related Quality of Life Program (HRQL) for the National Surgical Adjuvant Breast and Bowel Project (NSABP). Specific aims proposed for the CDA included: (a) Design and implementation of new HRQL components for planned NSABP treatment and prevention trials; (b) testing and implementation of data collection methods to be used in treatment and prevention trials; (c) analysis of HRQL data collected in the NSABP prevention and treatment trials; (d) refinement and extension of HRQL methods to analyze the data from new treatment and prevention studies; (e) enhancement of minority participation in NSABP trials. This is a final report on the work for this Career Development Award.

14. SUBJECT TERMS
breast cancer, health-related quality-of-life program, National Surgical Adjuvant Breast and Bowel Project

15. NUMBER OF PAGES
122

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Richard Day, Ph.D.
Department of Biostatistics
University of Pittsburgh

1. Introduction

This Career Development Award (CDA) was specifically intended to support Dr. Day in the development of a Health-Related Quality of Life Program (HRQL) for the National Surgical Adjuvant Breast and Bowel Project (NSABP). Specific aims proposed for the CDA included: (a) Design and implementation of new HRQL components for planned NSABP treatment and prevention trials; (b) testing and implementation of data collection methods to be used in treatment and prevention trials; (c) analysis of HRQL data collected in the NSABP prevention and treatment trials; (d) refinement and extension of HRQL methods to analyze the data from new treatment and prevention studies; (e) enhancement of minority participation in NSABP trials. This is a final report on the work completed under this CDA. The work completed over the period of Dr. Day's CDA will be summarized in terms of above stated aims.

2. Body

2.1 Design and implementation of new HRQL breast cancer components for planned NSABP treatment and prevention trials.

Summary of NSABP clinical trials protocols with Health-Related Quality of Life (HRQL) components designed and implemented as part of Dr. Day's CDA:

Treatment Trials:

a. **Protocol B23** - *Study to Evaluate the Effect on Quality of Life of Adriamycin Cyclophosphamide Therapy versus Cyclophosphamide, Methotrexate, and 5-Flourourciil Therapy in Women with Axillary Node-Negative, Estrogen-Receptor Negative Primary Invasive Breast Cancer*

b. **Protocol no. B-30** – *A Three Arm Randomized Trial to Compare Adjuvant Adriamycin and Cyclophosphamide Followed by Taxotere (AC-T); Adriamycin and Taxotere (AT); and Adriamycin, Taxotere and Cyclophosphamide (ATC) in Breast Cancer Patients with Positive Axillary Lymph Nodes.*

c. **Protocol no. B-32** – *A Randomized, Phase III Clinical Trial to Compare Sentinel Node Resection to Conventional Axillary Dissection in Clinically Node Negative Breast Cancer Patients*

d. **Protocol no. B-33** – *A Randomized, Placebo Controlled, Double-Blinded Trial Evaluating the Effect of Exemestane in Stage I and II Post-Menopausal Breast Cancer Patients Completing at least Five Years of Tamoxifen Therapy.*

e. **Protocol C-06** - A Clinical Trial Comparing Oral Uracil/Ftorafur (UFT) Plus Leucovorin (LV) With 5-Fluorouracil (5-FU) Plus LV In The Treatment Of Patients With Stages II And III Carcinoma Of The Colon

f. **Protocol C-07** - *Trial Comparing 5-Fluorourcil (5-FU) Plus Leucovorin(LV) and Oxaliplatin with 5-FU Plus LV for the Treatment of Patients with Stages II and III Carcinoma of the Colon.*

Prevention Trials:

a. **Protocol no. P-2** – *Study of Tamoxifen and Raloxifene (STAR).*

b. **Protocol STAR-Cog** - Effects of Selective Estrogen Receptor Modulators on Cognitive Aging: A Study of Tamoxifen, Raloxifene and Cognition.

2.2 Testing and implementation of data collection methods to be used in treatment and prevention trials

Operational Procedures to Reduce Missing and Delinquent HRQL Data – An

operational strategy for the reduction of missing and delinquent data was developed and implemented at NSABP during Dr. Day's CDA. Specific elements of this strategy included: (1) The use of missing data forms; (2) the inclusion of HRQL questionnaires in delinquency assessments; (3) periodic HRQL training sessions at national meetings; and, (4) the routine notification of study coordinators of scheduled HRQL examinations. Overall compliance rates for most HRQL studies remains at or above approximately 70%.

2.3 Analysis of HRQL Data Collected in the NSABP Prevention and Treatment Trials:

a. Peer Reviewed Papers (Appendix 1):

Day R, Ganz PA, Ganz, PA, Costantino JC., Cronin WM, Wickerham LA, Fisher B. Health-Related Quality of Life in and Tamoxifen in Breast Cancer Prevention: A Report from the NSABP Project P-1 Study. J Clin Oncology, 17, 1999, 2659-2669.

Day R, Ganz PA, Costantino JC. Tamoxifen and Depression: More Evidence from the NSABP's Breast Cancer Prevention (P-1) Randomized Study. JNCI, 93, 2001 (in press, 7 Nov. 2001 issue).

Day R, Quality of life and tamoxifen in breast cancer: a summary of the findings from the NSABP P-1 study. Annals of the New York Academy of Sciences (in press).

Land S, Wieand S, Day R, Have T, Costantino J, Lang W, Ganz P. Methodological issues in the analysis of quality of life data in clinical trials: illustrations from the NSABP Breast Cancer Prevention Program. In: M. Mesbah, B. Cole, M Lee (eds.), Statistical Design, Measurement and Analysis of Health Related Quality of Life. Klewler Academic Publishers (in press).

Kiebert G, Wait S, Bernhard J, Bezjak A, Cella D, Day R, Houghton J, Moinpiour C, Scott C, Stephens C. Practice and policy of measuring quality of life and health economics in cancer clinical trials: a survey among cooperative groups. Quality of Life Research 2000; 9(10):1073-80. (Appendix 4)

b. Submitted Papers:

Stephanie RL ,Kopec JA, Yothers G, Anderson S, Day R,Tang G, Ganz PA, Fisher B, Wolmark N., Health-Related Quality of Life in Axillary Node-Negative, Estrogen Receptor-Negative Breast Cancer Patients Undergoing AC versus CMF Chemotherapy: Findings from the National Adjuvant Breast and Bowel Project B-23. Submitted to the Journal of Clinical Oncology.

Day R, Cella D, Ganz PA, Daly MB, Rowland J, Wolter J. Determining the Feasibility and Usefulness of Microelectronic Adherence Monitoring Compared to Pill Counts

and Self-Reports in a Large, Multicenter Chemoprevention Trial. Submitted to Controlled Clinical Trials (in revision).

c. **Papers in Preperation:**

Chang CH, Cella D, Ganz PA, Day R. Scaling symptoms relevant when using hormonal therapies to prevent breast cancer: Results from the NSABP P-1 Study.

d. **Data Presentations and Posters:**

Day, R. Key Quality of Life Findings from the NSABP P-1 Breast Cancer Prevention Trial. Paper presented at NIH Workshop on Selective Estrogen Receptor Modulators (SERMs), April 26-28, 2000, Lister Hill Auditorium, NIH, Bethesda, MD.

Day, R. Development of an Integrated Program of Health-Related Quality of Life Research for the National Surgical Adjuvant Breast and Bowel Project. Poster presented Department of Defense, BCRP Era of Hope Meeting, June 8-11, 2000, Atlanta Hilton and Towers, Atlanta, GA.

Day, R. Does Tamoxifen Cause Depression? Paper presented at University of Pittsburgh, Graduate School of Public Health Lecture Series, May 12, 2000. University of Pittsburgh, Graduate School of Public Health, Pittsburgh, PA (Appendix 3).

Day, R. Initial HRQL Findings from the NSABP B-23 Protocol. Presentation at the NSABP National Meeting, June 12, 2000. New Orleans, LA.

Day, R. A Reviw of Health-Related Quality of Life Data from Phase III Clinical Trials of Fulvestrant and Other Hormonal Treatments
for Advanced Breast Cancer, Astra-Zeneca Workshop on Estrogen-Receptor Downregulation. March 9-10, 2002 Sanibel Island, FL.

2.4 Refinement and extension of HRQL methods to analyze the data from new treatment and prevention studies

The statistical and methodological techniques developed during this CDA were utilized within the specific data analyses summarized in Section 2.3.a.:

Improved longitudinal methods for the investigation of changes in reported quality of life were published in:

Day R, Ganz PA, Ganz, PA, Costantino JC., Cronin WM, Wickerham LA, Fisher B. Health-Related Quality of Life in and Tamoxifen in Breast Cancer Prevention: A Report from the NSABP Project P-1 Study. J Clin Oncology, 17, 1999, 2659-2669.

Day R, Ganz PA, Costantino JC. Tamoxifen and Depression: More Evidence from the NSABP's Breast Cancer Prevention (P-1) Randomized Study. JNCI, 93, 2001 (in press, 7 Nov. 2001 issue).

Methodological developments in the imputation and estimation of missing data were discussed and published in:

Day R, Ganz PA, Costantino JC. Tamoxifen and Depression: More Evidence from the NSABP's Breast Cancer Prevention (P-1) Randomized Study. JNCI, 93, 2001 (in press, 7 Nov. 2001 issue).

Stephanie RL ,Kopec JA, Yothers G, Anderson S, Day R,Tang G, Ganz PA, Fisher B, Wolmark N., Health-Related Quality of Life in Axillary Node-Negative, Estrogen Receptor-Negative Breast Cancer Patients Undergoing AC versus CMF Chemotherapy: Findings from the National Adjuvant Breast and Bowel Project B-23. Submitted to the Journal of Clinical Oncology.

Land S, Wieand S, Day R, Have T, Costantino J, Lang W, Ganz P. Methodological issues in the analysis of quality of life data in clinical trials: illustrations from the NSABP Breast Cancer Prevention Program. In: M. Mesbah, B. Cole, M Lee (eds.), Statistical Design, Measurement and Analysis of Health Related Quality of Life. Klewler Academic Publishers (in press).

2.5 Enhancement of minority participation in NSABP trials and the implementation of measures focusing on HRQL-related issues in women of color

Limited progress was made on this key aim over the course of Dr. Day's CDA. Collaborative work was carried out with NSABP HQ staff to insure participation of minority women in the P-2, but no overall program focusing on increasing the numbers of women of color participating in NSABP HRQOL trials could be implemented. This objective also became part of the mandate of the new Behavioral and Health Outcomes Committee which was established in 2001 under the leadership of Dr. Patricia A Ganz.

3. Key Research Accomplishments

- **Eight new NSABP treatment trials containing a Health-Related Quality of Life component designed and implemented.**

- **Development of the current NSABP operational system for the reduction of missing and delinquent data in NSABP HRQOL trials.**

- **Publication of five major peer-reviewed papers on NSABP HRQOL data, with three additional manuscripts under submission to peer reviewed journals.**

- **Development of Behavior and Heath Outcomes Committee to steer future HRQOL work at NSABP.**

4. Reportable Outcomes

a. Peer Reviewed Papers:

Day R, Ganz PA, Ganz, PA, Costantino JC., Cronin WM, Wickerham LA, Fisher B. Health-Related Quality of Life in and Tamoxifen in Breast Cancer Prevention: A Report from the NSABP Project P-1 Study. J Clin Oncology, 17, 1999, 2659-2669.

Day R, Ganz PA, Costantino JC. Tamoxifen and Depression: More Evidence from the NSABP's Breast Cancer Prevention (P-1) Randomized Study. JNCI, 93, 2001 (in press, 7 Nov. 2001 issue).

Day R, Quality of life and tamoxifen in breast cancer: a summary of the findings from the NSABP P-1 study. Annals of the New York Academy of Sciences (in press).

Land S, Wieand S, Day R, Have T, Costantino J, Lang W, Ganz P. Methodological issues in the analysis of quality of life data in clinical trials: illustrations from the NSABP Breast Cancer Prevention Program. In: M. Mesbah, B. Cole, M Lee (eds.), Statistical Design, Measurement and Analysis of Health Related Quality of Life. Klewler Academic Publishers (in press).

Kiebert G, Wait S, Bernhard J, Bezjak A, Cella D, Day R, Houghton J, Moinpiour C, Scott C, Stephens C. Practice and policy of measuring quality of life and health economics in cancer clinical trials: a survey among cooperative groups. Quality of Life Research 2000; 9(10):1073-80. (Appendix 4)

b. Submitted Papers:

Stephanie RL ,Kopec JA, Yothers G, Anderson S, Day R,Tang G, Ganz PA, Fisher B, Wolmark N., Health-Related Quality of Life in Axillary Node-Negative, Estrogen Receptor-Negative Breast Cancer Patients Undergoing AC versus CMF Chemotherapy: Findings from the National Adjuvant Breast and Bowel Project B-23. Submitted to the Journal of Clinical Oncology.

c. Data Presentations and Posters:

Day, R. Key Quality of Life Findings from the NSABP P-1 Breast Cancer Prevention Trial. Paper presented at NIH Workshop on Selective Estrogen Receptor Modulators (SERMs), April 26-28, 2000, Lister Hill Auditorium, NIH, Bethesda, MD.

Day, R. Development of an Integrated Program of Health-Related Quality of Life Research for the National Surgical Adjuvant Breast and Bowel Project. Poster presented Department of Defense, BCRP Era of Hope Meeting, June 8-11, 2000, Atlanta Hilton and Towers, Atlanta, GA.

Day, R. Does Tamoxifen Cause Depression? Paper presented at University of Pittsburgh, Graduate School of Public Health Lecture Series, May 12, 2000. University of Pittsburgh, Graduate School of Public Health, Pittsburgh, PA (Appendix 3).

Day, R. Initial HRQL Findings from the NSABP B-23 Protocol. Presentation at the NSABP National Meeting, June 12, 2000. New Orleans, LA.

Day, R. A Reviw of Health-Related Quality of Life Data from Phase III Clinical Trials of Fulvestrant and Other Hormonal Treatments
for Advanced Breast Cancer, Astra-Zeneca Workshop on Estrogen-Receptor Downregulation. March 9-10, 2002 Sanibel Island, FL.

Conclusion

When this Career Development Award was initiated, NSABP had little or no research activities in the area of Health-Related Quality of Life (HRQOL) related to its treatment trials. Over the course of this CDA, eight new NSAPB treatment protocols have been designed and implemented that have a significant HRQOL component. In addition, a complete operational structure has been designed and implemented for the handling of HRQOL data and for the reduction of delinquent and missing data. During this period, eight collaborative HRQOL papers have been written on NSABP data and five have been published in peer reviewed journals. Recently, a new Behavior and Health Outcome Committee has been established which will have responsibility for directing future NSABP work in the HRQOL area. In summary, all of the major objectives outlined in the original CDA application have been successfully completed, with the exception of the further expansion of participation of women of color in NSABP HRQOL studies. Hopefully, this last objective will be successfully addressed in future work.

Personnel Receiving pay from this grant:

Richard Day, Ph.D., Department of Biostatisitcis, University of Pittsburgh
Lisa Weissfield, Ph.D., Department of Biostatistics, University of Pittsburgh

6 References

APPENDIX 1

Publications and Manuscripts

Publications in order of inclusion:

B. Health-Related Quality of Life in and Tamoxifen in Breast Cancer Prevention: A Report from the NSABP Project P-1 Study. J Clin Oncology, 17, 1999, 2659-2669.

Day R, Ganz PA, Costantino JC. Tamoxifen and Depression: More Evidence from the NSABP's Breast Cancer Prevention (P-1) Randomized Study. JNCI, 93, 2001 (in press, 7 Nov. 2001 issue).

Day R, Quality of life and tamoxifen in breast cancer: a summary of the findings from the NSABP P-1 study. Annals of the New York Academy of Sciences (in press).

Land S, Wieand S, Day R, Have T, Costantino J, Lang W, Ganz P. Methodological issues in the analysis of quality of life data in clinical trials: illustrations from the NSABP Breast Cancer Prevention Program. In: M. Mesbah, B. Cole, M Lee (eds.), Statistical Design, Measurement and Analysis of Health Related Quality of Life. Klewler Academic Publishers (in press).

Kiebert G, Wait S, Bernhard J, Bezjak A, Cella D, Day R, Houghton J, Moinpiour C, Scott C, Stephens C. Practice and policy of measuring quality of life and health economics in cancer clinical trials: a survey among cooperative groups. Quality of Life Research 2000; 9(10):1073-80. (Appendix 4)

Stephanie RL ,Kopec JA, Yothers G, Anderson S, Day R,Tang G, Ganz PA, Fisher B, Wolmark N., Health-Related Quality of Life in Axillary Node-Negative, Estrogen Receptor-Negative Breast Cancer Patients Undergoing AC versus CMF Chemotherapy: Findings from the National Adjuvant Breast and Bowel Project B-23. Submitted to the Journal of Clinical Oncology.

Day R, Cella D, Ganz PA, Daly MB, Rowland J, Wolter J. Determining the Feasibility and Usefulness of Microelectronic Adherence Monitoring Compared to Pill Counts and Self-Reports in a Large, Multicenter Chemoprevention Trial. Submitted to Controlled Clinical Trials (in revision).

Chang CH, Cella D, Ganz PA, Day R. Scaling symptoms relevant when using hormonal therapies to prevent breast cancer: Results from the NSABP P-1 Study.

Health-Related Quality of Life and Tamoxifen in Breast Cancer Prevention: A Report From the National Surgical Adjuvant Breast and Bowel Project P-1 Study

By Richard Day, Patricia A. Ganz, Joseph P. Costantino, Walter M. Cronin, D. Lawrence Wickerham, and Bernard Fisher

Purpose: **This is the initial report from the health-related quality of life (HRQL) component of the National Surgical Adjuvant Breast and Bowel Project Breast Cancer Prevention Trial. This report provides an overview of HRQL findings, comparing tamoxifen and placebo groups, and advice to clinicians counseling women about the use of tamoxifen in a prevention setting.**

Patients and Methods: **This report covers the baseline and the first 36 months of follow-up data on 11,064 women recruited over the first 24 months of the study. Findings are presented from the Center for Epidemiological Studies-Depression Scale (CES-D), the Medical Outcomes Study 36-Item Short Form Health Status Survey (MOS SF-36) and sexual functioning scale, and a symptom checklist.**

Results: **No differences were found between placebo and tamoxifen groups for the proportion of participants scoring above a clinically significant level on the CES-D. No differences were found between groups for the MOS SF-36 summary physical and mental scores. The mean number of symptoms reported was consistently higher in the tamoxifen group and was associated with vasomotor and gynecologic symptoms. Significant increases were found in the proportion of women on tamoxifen reporting problems of sexual functioning at a definite or serious level, although overall rates of sexual activity remained similar.**

Conclusion: **Women need to be informed of the increased frequency of vasomotor and gynecologic symptoms and problems of sexual functioning associated with tamoxifen use. Weight gain and depression, two clinical problems anecdotally associated with tamoxifen treatment, were not increased in frequency in this trial in healthy women, which is good news that also needs to be communicated.**

J Clin Oncol 17:2659-2669.

THIS IS THE INITIAL report of the findings from the health-related quality of life (HRQL) component of the National Surgical Adjuvant Breast and Bowel Project (NSABP) Breast Cancer Prevention Trial (P-1), a multicenter, double-blinded, placebo-controlled clinical trial. The purpose of this report is to provide a concise overview of the P-1 HRQL findings and an assessment of the effects of tamoxifen, when used as a preventative agent, on self-reported symptoms and everyday physical, emotional, and social functioning. Recommendations have been provided that may be helpful to physicians involved in counseling women considering the use of tamoxifen in the setting of prevention.

The primary objective of the P-1 study was to evaluate whether 5 years of tamoxifen therapy would reduce the incidence of invasive breast cancer in women at an increased risk for the disease. Secondary objectives were to assess the incidence of ischemic heart disease, bone fractures, and other events, such as depression, that might be associated with the use of tamoxifen. Eligible participants were randomized either to 20 mg daily of tamoxifen or to a placebo for a planned 5 years.

Detailed descriptions of the rationale, planning, and design of the of the Breast Cancer Prevention Trial and the HRQL component of the P-1 study, as well as specific instruments, have been provided in separate reports.[1-3]

PATIENTS AND METHODS

Participant Cohort and HRQL Data

This report covers the baseline HRQL examination and the first 36 months of follow-up data on 11,064 women recruited over the first 24 months (June 1, 1992, to May 31, 1994) of the study. This cohort of women represents 82.6% of the total P-1 accrual (n = 13,388). Restrictions were imposed on the initial HRQL report for two reasons. First, by limiting our attention to this cohort of women, we avoided the potential bias created by events beginning in March 1994,[4,5] which resulted in a suspension of accrual to the P-1 study. Second, a focus on the first 36 months of data collection permitted improved control over types of missing HRQL data because all 11,064 participants should have completed the eight scheduled examinations before the disclosure of the results of the trial in the spring of 1998.

From the National Surgical Adjuvant Breast and Bowel Project (NSABP) Operations and Biostatistical Centers, Pittsburgh, PA, and Jonsson Comprehensive Cancer Center, University of California Los Angeles, Los Angeles, CA.

Submitted December 7, 1998; accepted April 22, 1999.

Supported by public health service grants from the National Cancer Institute (NCI-U10-CA-37377/69974) and a career development award from the Department of Defense (DAMD17-97-1-7058).

Address reprint requests to Richard Day, PhD, Department of Biostatistics, Graduate School of Public Health, 130 DeSoto St, University of Pittsburgh, Pittsburgh, PA 15261; email rdfac@vms.cis.pitt.edu.

0732-183X/99/1709-2659

Instruments

The 104-item P-1 HRQL Questionnaire[3] was composed of the Center for Epidemiological Studies–Depression Scale (CES-D, 20 items), the Medical Outcomes Study (MOS) 36-Item Short Form Health Status Survey (SF-36, 36 items), the MOS sexual functioning scale (five items), and a symptom checklist (SCL, 43 items). The questionnaire was scheduled to be administered to all participants before randomization (baseline), at 3 months, at each succeeding 6-month examination for the planned 5 years of treatment, and for 1 year after treatment was completed.

Data Completeness

The P-1 study has multiple, complex levels of missing and incomplete data. In the case of self-administered instruments, such as the HRQL questionnaire, participants could leave items blank by error or because they did not wish to answer the question.[6] Beyond this, the staffs of collaborating centers were generally unable to collect self-administered instruments on participants who quit taking pills because they no longer appeared for follow-up examinations, although many of these participants can still be observed for primary end points (eg, breast cancer and fractures). In addition, there are participants who did not complete all of the scheduled follow-up HRQL questionnaires because of the disclosure of the trial results in the spring of 1998,[1] although they are still observed for primary end points. Finally, a small proportion of participants (1.7%) were lost to follow-up, even for primary end points.

Statistical Analysis

The P-1 HRQL data set is composed of multiple HRQL instruments, each with its own psychometric properties and research history.[3] This complexity is magnified by the fact that data distributions and patterns of missing data differ across the various instruments included in the HRQL questionnaire. In addition, sample sizes are large, resulting in the possibility of statistically significant findings for clinically negligible effects. All of these considerations argue for future detailed analyses of the data from each specific instrument. In this initial report, however, our aims were essentially descriptive in nature and emphasized basic comparisons of the two trial groups. In making these comparisons, we seek to identify consistent differences, between the trial groups, using simple nonparametric procedures. The sign test[7] is used to examine the consistency of binary differences ($\pm$) between the two trial groups across time, independent of the magnitude of these differences. A one-sided alternative is routinely used because tamoxifen is expected to have a negative effect on most short-term measures of HRQL. Friedman's test,[7] implemented as a generalization of the paired sign test,[8] was used as a nonparametric analog to the two-way analysis of variance when we wanted to block on a specific factor, such as age group. Positive findings, with regard to consistent differences between trial groups, were independently reviewed for magnitude to assess their clinical and functional significance for the participants' quality of life.

Clinical experience, as well as initial statistical investigations of the P-1 HRQL data set, suggested that the age of the study participants was a key factor contributing to the observed distribution of HRQL measures. Hence, the results presented here from various HRQL instruments were routinely stratified by three age groups (35 to 49 years, 50 to 59 years, and 60 years or older) that generally paralleled menopausal status. Relative risks (RRs) or absolute differences in mean counts are presented in the tables to estimate differences in effect size between the two groups.

Imputation procedures for missing items in otherwise complete scales were only used for eight SF-36 subscales, as recommended in the SF-36 scoring manual.[9] No data imputation was carried out for other scales, and incomplete scales were considered missing.[6]

RESULTS

Table 1 lists the demographic, medical, and behavioral characteristics of our participant cohort of 11,064 women by trial group. These data show that the women in the P-1 study were predominately white (96%), well educated (65% $\geq$ some college), married (70%), professional and technically trained (68.2%), currently employed (64.9%), and reported a middle- to upper-middle class family income (median, $35,000 to $49,999). None of the variables in Table 1 show a striking imbalance between the two trial groups.

Figure 1 charts the overall proportion and total numbers of women completing the HRQL questionnaire at each examination. It provides a general measure of comparative participant adherence with regard to the HRQL questionnaire in the two trial groups. Both trial groups showed a consistent decline in HRQL adherence across the first 36 months of the study, averaging 4.2% per examination in the placebo group and 4.6% per examination in the tamoxifen group. The proportion of HRQL-adherent participants was smaller in the tamoxifen than in the placebo group at every one of the seven follow-up examinations (sign test, $P = .0078$), with a maximum difference of 3.1% occurring at 36 months.

A number of demographic, clinical, and HRQL variables were examined to investigate whether differences could be detected between the women who failed to complete the HRQL questionnaire at 36 months in the tamoxifen and the placebo groups. These variables included mean age (tamoxifen = 53.1 years *v* placebo = 53.5 years) and mean RR (5.42 *v* 5.43), treatment status (10.1% *v* 10.5% on treatment), breast cancer in a first-degree relative (76.89% *v* 78.40%), prior estrogen use (32.5% *v* 33.3%), mean maximum CES-D score (12.52 *v* 12.46), and mean maximum number of reported symptoms on the SCL (14.2 *v* 13.9). These comparisons suggested that participants who failed to complete the HRQL questionnaire in each group were similar cohorts of women.

When, within a treatment group, the same variables were used to compare HRQL adherent and nonadherent women, only the treatment status variable was different between the two groups. A significantly greater proportion of HRQL-adherent women in both groups remained on treatment (87.0% *v* 89.6%) compared with HRQL-nonadherent women (10.1% *v* 10.5%). In other words, adherence in the HRQL component of P-1 was largely a reflection of treatment adherence. This was because most collaborating centers did not have the staff resources to administer the HRQL

Table 1. Demographic, Clinical, and Health Behavior Characteristics of P-1 HRQL Study Participants (N = 11,064)

Characteristic	Placebo		Tamoxifen		Total	
	No. of Patients	%	No. of Patients	%	No. of Patients	%
Age, years						
Mean ± SD	53.83 ± 9.167		53.82 ± 9.184		53.83 ± 9.175	
Median	52		52		52	
Range	35-79		35-78		35-79	
Ethnicity						
White	5,290	95.54	5,282	95.57	10,572	95.55
Hispanic	63	1.14	49	0.89	112	1.01
Black	88	1.59	95	1.72	183	1.65
Asian	35	0.63	37	0.67	72	0.65
Other	47	0.84	39	0.71	86	0.78
Missing	14	0.25	25	0.45	39	0.35
Education						
Grade school	61	1.10	66	1.19	127	1.15
Some high school	248	4.48	218	3.94	466	4.21
High school graduate	1,003	18.11	1,009	18.26	2,012	18.19
Vocational school	593	10.71	614	11.11	1,207	10.91
Some college	1,180	21.31	1,194	21.60	2,374	21.46
Associate degree	349	6.30	349	6.31	698	6.31
College graduate	664	11.99	732	13.24	1,396	12.62
Professional school	546	9.86	519	9.39	1,065	9.63
Master's degree	726	13.11	684	12.38	1,410	12.74
Doctoral degree	133	2.40	106	1.92	239	2.16
Missing	34	0.61	36	0.65	70	0.63
Employment						
Unemployed	239	4.32	229	4.14	468	4.23
Retired	925	16.71	938	16.97	1,863	16.84
Full-time homemaker	660	11.92	670	12.12	1,330	12.02
Student	30	0.54	33	0.60	63	0.57
Employed full-time	2,713	49.00	2,682	48.53	5,395	48.76
Employed part-time	880	15.89	878	15.89	1,758	15.89
On medical leave	25	0.45	24	0.43	49	0.44
Permanently disabled	51	0.92	47	0.85	98	0.89
Missing	14	0.25	26	0.47	40	0.36
Occupation						
Homemaker	849	15.33	843	15.25	1,692	15.29
Professional	2,207	39.86	2,188	39.59	4,395	39.72
Technical	1,573	28.41	1,548	28.01	3,121	28.21
Services	487	8.80	487	8.81	974	8.80
Operators	92	1.66	94	1.70	186	1.68
Other	315	5.69	341	6.17	656	5.93
Missing	14	0.25	26	0.47	40	0.36
Income						
Under $10,000	211	3.81	161	2.91	372	3.36
$10,000-$19,999	549	9.91	571	10.33	1,120	10.12
$20,999-$34,999	1,127	21.35	1,170	21.17	2,297	20.76
$35,000-$49,999	936	16.90	984	17.80	1,920	17.35
$50,000-$74,999	1,153	20.82	1,151	20.83	2,304	20.82
$75,000-$99,000	511	9.23	478	8.65	989	8.94
$100,000 or more	564	10.19	521	9.43	1,085	9.81
Unanswered	296	5.35	301	5.45	597	5.40
Missing	190	3.43	190	3.44	380	3.43
RR of breast cancer						
1-2	416	7.51	416	7.53	832	7.52
2-3	929	16.78	865	15.65	1,794	16.21
3-5	2,074	37.46	2,154	38.97	4,228	38.21
5-10	1,618	29.22	1,605	29.04	3,223	29.13
10+	500	9.03	487	8.81	987	8.92
1st degree relatives w/breast cancer						
0	1,238	22.36	1,191	21.56	2,429	21.95
1	3,239	58.50	3,250	58.80	6,489	58.65
2	903	16.31	902	16.32	1,805	16.31
≥ 3	157	2.83	184	3.32	341	3.09
Marital status						
Never married	398	7.19	394	7.13	792	7.16
Presently married	3,843	69.41	3,876	70.43	7,719	69.77
Marriage-like	139	2.51	125	2.26	264	2.39
Divorced	748	13.51	707	12.79	1,455	13.15
Widowed	395	7.13	399	7.22	794	7.18
Unknown	0	0	1	0.02	1	0.01
Missing	14	0.25	25	0.45	39	0.35
Smoking						
Smoked at least 100 cigarettes in lifetime	2,697	48.83	2,729	49.60	5,470	50.39
Smoked at least 100 cigarettes in lifetime and currently smoke	705	12.76	712	12.94	1,417	12.85
Alcohol						
Never use	1,138	20.60	1,128	20.50	2,266	20.55
Some days	4,129	74.76	4,147	75.37	8,276	75.07
Every day	256	4.64	227	4.13	483	4.38
Previous estrogen use	1,171	31.98	1,838	33.25	3,609	32.62
Both ovaries removed	797	14.39	813	14.71	1,610	14.55
Menstrual period stopped	3,658	66.06	3,685	66.67	7,343	66.37

questionnaire via the telephone or mail to women who stopped treatment and failed to appear for their scheduled follow-up visits.

By the 36-month examination, 3,421 women had stopped their assigned treatment and failed to fill out the HRQL questionnaire for at least 6 months. Table 2 lists the primary reasons these women gave for stopping treatment. The placebo and tamoxifen groups did not differ with regard to protocol-specified events, such as invasive breast cancer, depression, or deep vein thrombosis, or other medical reasons, such as anxiety disorders or cardiovascular conditions. Hot flashes were clearly the most frequently reported sign or symptom that caused women to stop their assigned treatment (251 women); they occurred most often in the tamoxifen group (184 women). When stopping their as-

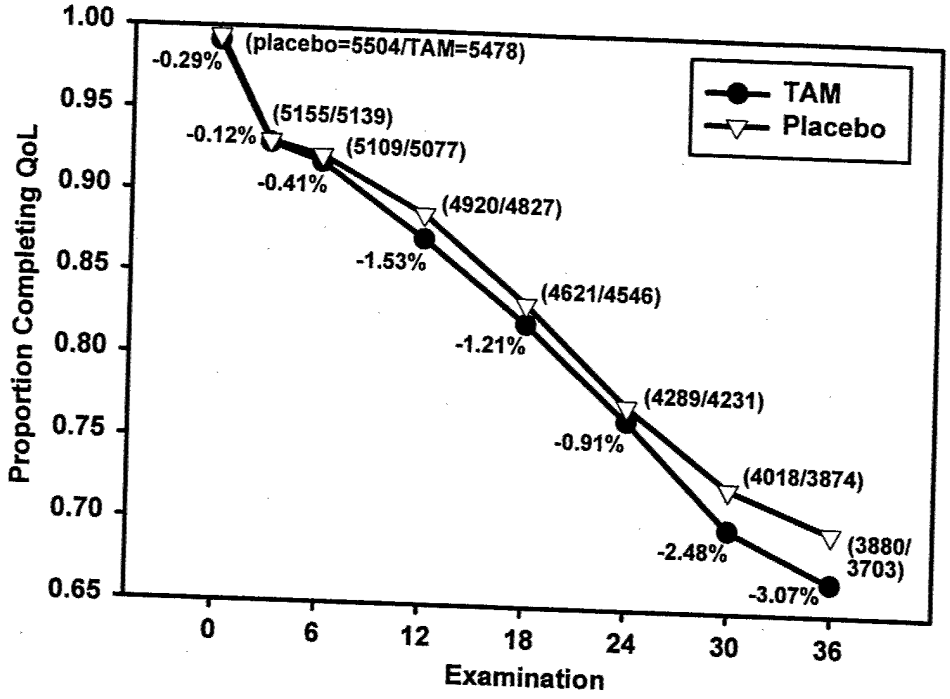


Fig 1. Proportion of participants in the tamoxifen group and placebo group completing HRQL questionnaire by examination (placebo, n = 5,537; tamoxifen, n = 5,527). Figures on chart are the number of women in the placebo/tamoxifen groups completing the HRQL questionnaire and the difference between TAM and placebo groups in terms of percent missing HRQL data.

signed treatment, participants in the placebo group were more likely to cite other nonmedical reasons, such as fear of side effects, change of mind, or desire to adopt an alternative therapy (eg, hormone replacement).

Table 3 shows the proportion of P-1 participants, by age group and examination, who scored above the most frequently used clinical cutoff (≥ 16) on the CES-D.[10,11] The youngest age group (35 to 49 years) in both trial groups consistently had the highest proportion of members scoring above the clinical cutoff, followed by the 50- to 59-year-old age group (Friedman test, $P = .001$ tamoxifen and placebo). The RRs listed in Table 3 show that, for all three age groups, the magnitude of the differences is small, and there was no consistent excess of participants in the tamoxifen group scoring above the clinical cutoff on the CES-D when compared with the placebo group. Similar findings with regard to the relationship between the two trial groups emerged from the analysis of the five-item mental health subscale on the MOS SF-36 (not shown).

The results of the SF-36 are summarized using the physical component summary (PCS) and mental component summary (MCS) scores[12] and the eight SF-36 subscales. The PCS and MCS scores represent aggregate measures that combine data from the eight subscales generally reported on the SF-36. The PCS aggregates data from the Physical Functioning, Role-Physical, Bodily Pain, and General Health subscales, while the MCS draws on data from the Vitality, Social Functioning, Role-Emotional, and Mental Health subscales. The PCS and MCS are scored using norm-based

Table 2. Reasons for Stopping Assigned Therapy by Participants Not Completing Quality of Life Questionnaire (Baseline to 36-Month Examination, n = 3421)

Reason for Stopping Assigned Therapy	Tamoxifen		Placebo		Total	
	No. of Patients	%	No. of Patients	%	No. of Patients	%
Protocol specified event	164	9.1	154	9.6	318	9.3
Reported signs or symptoms	545	30.2	336	20.8	881	25.8
Other medical	342	18.9	280	17.3	622	18.2
Other nonmedical	753	41.7	842	52.1	1595	46.6
Unknown	2	0.1	3	0.2	5	0.1
Total	1806	52.8	1615	47.2	3421	100.0

Table 3. Proportion of Participants in Tamoxifen Arm With a Clinically Significant Score (≥ 16) on the CES-D by Age Group and Examination

Examination	Age Group							
	35-49 Years		50-59 Years		≥ 60 Years		Overall	
	TAM	RR*	TAM	RR*	TAM	RR*	TAM	RR*
Baseline	0.074	1.03	0.082	1.28	0.058	0.918	0.071	1.07
3 months	0.122	1.10	0.104	1.05	0.085	1.08	0.105	1.08
6 months	0.138	1.06	0.114	1.00	0.093	0.910	0.117	1.00
12 months	0.128	0.937	0.122	0.999	0.096	0.989	0.116	0.968
18 months	0.139	0.892	0.126	0.918	0.101	0.929	0.123	0.908
24 months	0.143	1.02	0.124	0.980	0.095	0.924	0.122	0.980
30 months	0.142	0.978	0.107	0.961	0.104	0.934	0.120	0.959
36 months	0.135	0.898	0.111	1.04	0.097	0.887	0.116	0.930

Abbreviation: TAM, tamoxifen.
*RR = TAM/placebo.

methods; both component scores have a mean of 50 and a SD of 10 in the general United States (U.S.) population. This means that the PCS and MCS can be meaningfully compared with one another, and their scores have a direct interpretation in relation to the distribution of scores in the general U.S. population.

Figure 2 charts the PCS and MCS for the tamoxifen and placebo groups at each examination and by age group. As expected, mean PCS declines across the age groups. At follow-up examinations, the tamoxifen group was consistently lower on the PCS only in the 50- to 59-year-old age group (one-sided sign test, $P = .065$). However, the absolute differences were small, approximating one tenth of an SD. With regard to the MCS, all of the age groups scored above the mean MCS for the general U.S. population, and no consistent differences emerged between the two trial groups. Figure 3 summarizes the overall data from eight subscales on which the component subscores are based.

Table 4 lists the mean number of symptoms reported on the 43-item SCL by age group and examination. The mean number of symptoms reported was consistently highest in the 50- to 59-year-old age group, followed by the 35- to 49-year-old and 60 years or older age groups (Friedman test, $P = .001$ tamoxifen and placebo). The participants in the tamoxifen group also reported a small but consistent excess in the mean number of symptoms (< one) reported at 19 of the 21 age-stratified follow-up examinations (3 to 36 months; one-sided sign test, 35 to 49 years, $P = .0078$; 50 to 59 years and ≥ 60 years, $P = .065$) (Table 4).

Table 5 provides information on the proportion of women in the tamoxifen and placebo groups who reported symptoms on the SCL at least once during the treatment period, ie, the period excluding baseline but including the seven follow-up examinations. The five symptoms with the greatest relative difference between the two trial groups are given for each age group, and the 10 symptoms with the greatest relative difference are presented for all participants combined.

Tables 6 and 7 give detailed information, by age group and examination, on the reported frequency of hot flashes and vaginal discharge in the trial groups. The proportion of participants who reported hot flashes was elevated in all age groups of the tamoxifen group at every follow-up examination. Among the participants in the tamoxifen group, the 50- to 59-year-old age group had the largest proportion of women reporting hot flashes at each examination (median, 69.8%; Friedman test, $P = .001$), but the youngest age group (35 to 49 years) showed the greatest relative increase in proportion of women reporting hot flashes (median RR, 1.50; Friedman test, $P = .011$). Vaginal discharge was the most consistently elevated symptom in the tamoxifen group. The youngest age group (35 to 49 years) had the greatest proportion of participants reporting vaginal discharge at each examination (median, 35.5%; Friedman test, $P < .001$), and the oldest age group (≥ 60 years) reported the greatest increase of vaginal discharge relative to the placebo controls (median RR, 3.05; Friedman test, $P = .005$).

Figure 4 summarizes the information from the five items on the MOS sexual functioning scale. Figure 4A shows that a greater proportion of participants in the tamoxifen group, as compared with the placebo group, reported being sexually active during the 6 months before each follow-up examination. Although apparently consistent ($P = .031$), the absolute difference was small (mean, 0.78%) and may have been caused by chance. Figure 4B through 4E show that a small but consistently larger percentage of participants in the tamoxifen group reported a definite or serious problem in three of the four specific domains of sexual functioning during the follow-up period.

DISCUSSION

We observed in our earlier article[3] that measuring the impact of new treatments on HRQL is particularly important within the context of disease-prevention and health-promotion trials. Compared with patients suffering from clinically manifest disease, decrements in overall quality of life are likely to have a much greater impact on the subjective appraisal of treatment acceptability and the maintenance of long-term treatment adherence among high-risk but otherwise healthy individuals. This report covers the initial HRQL findings from a large, multicenter chemoprevention trial, which has shown that tamoxifen reduced the risk of invasive breast cancer in high-risk women by 49% during the first 5 years of administration. Given the apparent clinical efficacy of tamoxifen in the prevention setting, it is important to assess whether the various secondary effects of the drug might act to reduce this practical efficacy.[13-15]

The cohort of women taking part in the P-1 study clearly was not representative of the general population. They were predominately white, well educated, and middle class, with a strong professional and technical orientation. The initial HRQL findings presented in this report must be assessed within the context of the socioeconomic and cultural characteristics of the P-1 study cohort.

The subcohort of women discussed in this report represent 82.6% of the total study cohort. This subcohort was chosen to exclude potential biases, because of external factors eventuating in the suspension of accrual in P-1, and to control for the amount and types of missing data. Despite this, we still lost 31.5% of our participants by the 36-month follow-up examination. This proportion closely approximates the 10%-per-year loss to follow-up rate predicted at

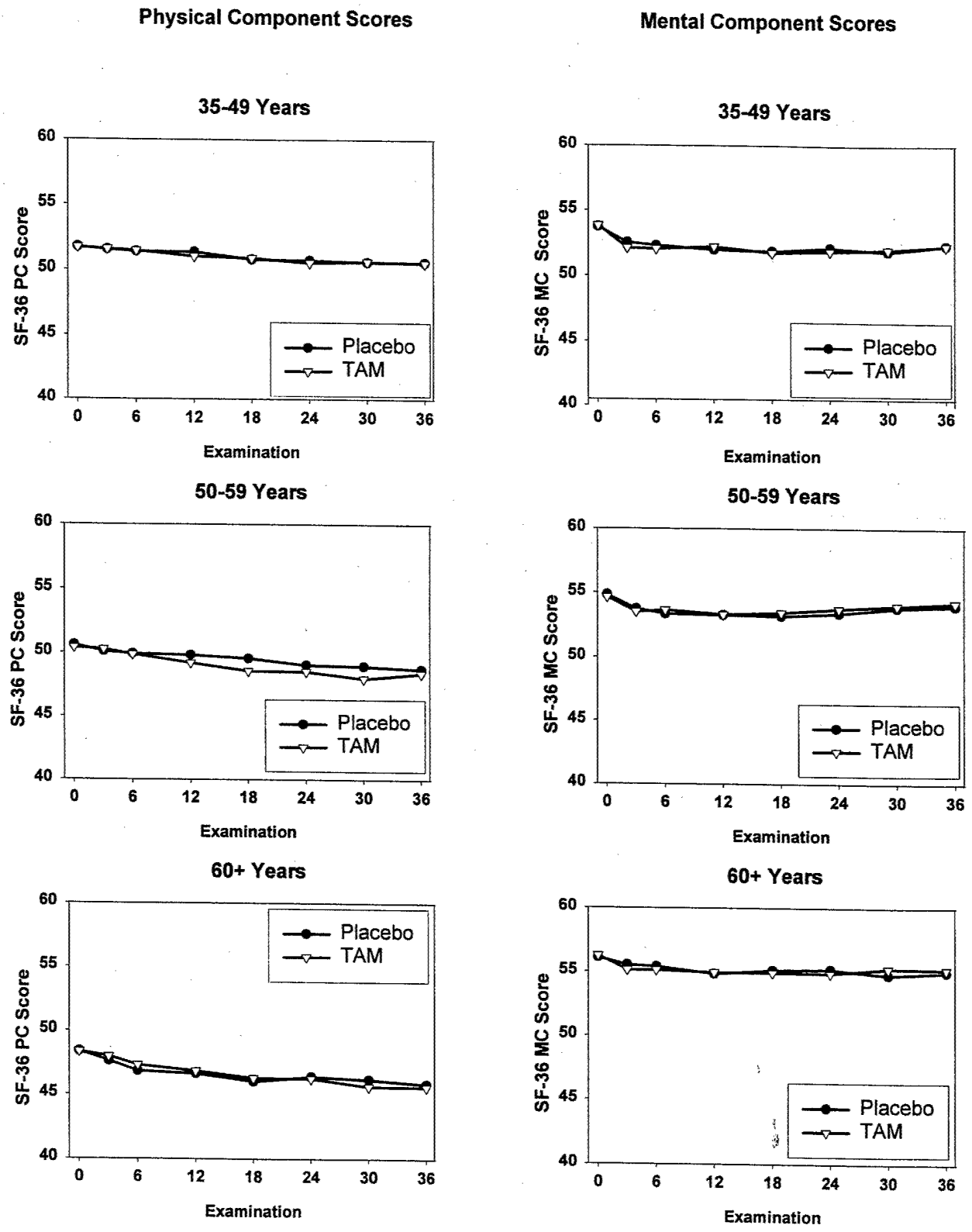


Fig 2. Mean scores by age group and examination on SF-36 physical and mental component scores (higher scores represent better quality of life).

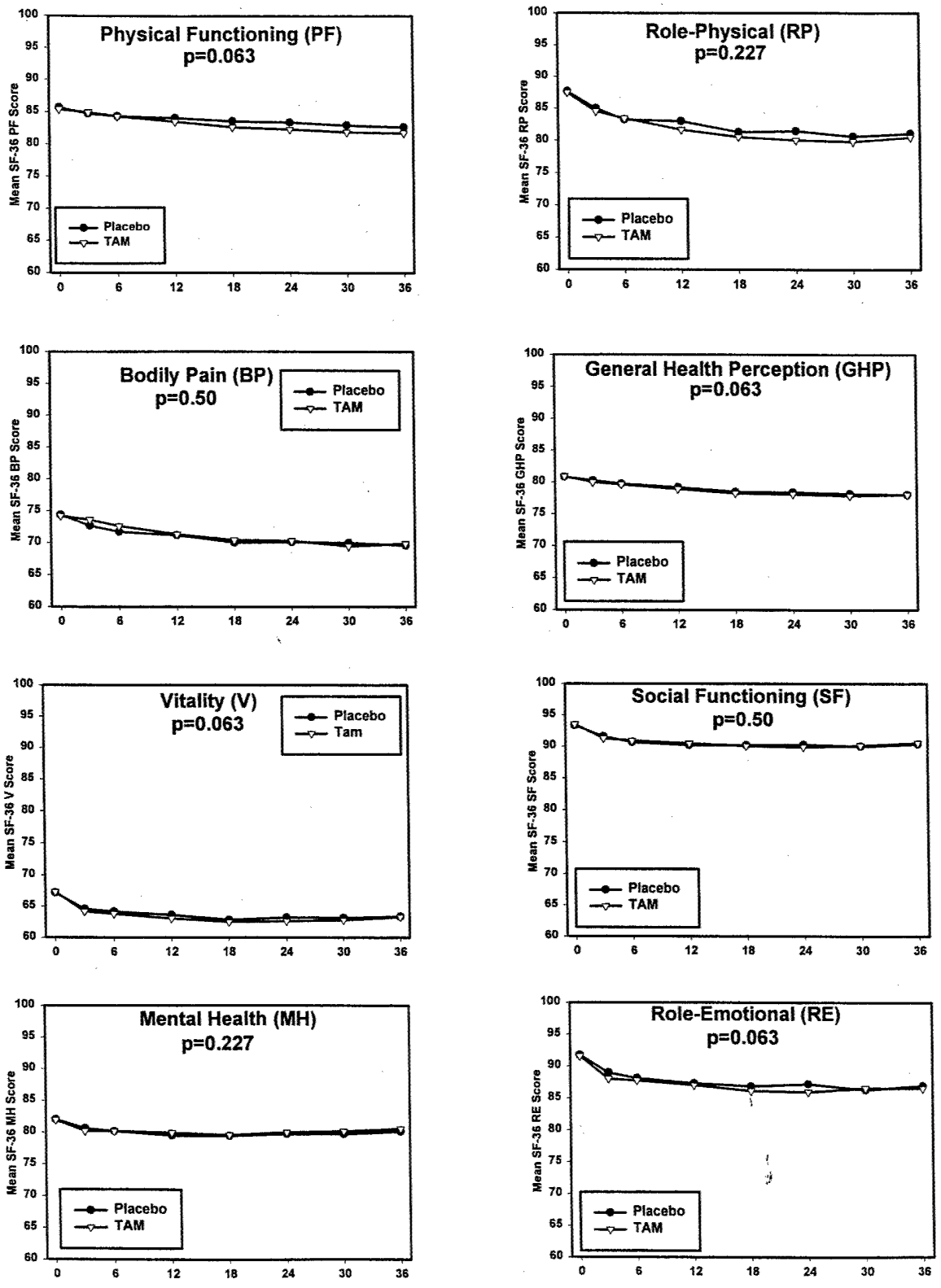


Fig 3. Mean SF-36 subscale scores by examination.

Table 4. Mean Number of Total Symptoms Reported on Symptom Checklist by Age Group and Examination

	Age Group							
	35-49 Years		50-59 Years		≥ 60 Years		Overall	
Examination	TAM	Difference*	TAM	Difference*	TAM	Difference*	TAM	Difference*
Baseline	8.84	+ 0.114	9.76	+ 0.236	8.89	− 0.030	9.14	+ 0.110
3 months	9.96	+ 0.319	10.54	− 0.006	9.63	− 0.166	10.04	+ 0.077
6 months	10.43	+ 0.564	11.06	+ 0.304	10.06	+ 0.011	10.51	+ 0.322
12 months	10.87	+ 0.521	11.54	+ 0.655	10.43	+ 0.076	10.95	+ 0.429
18 months	11.08	+ 0.614	11.51	+ 0.452	10.65	+ 0.292	11.08	+ 0.469
24 months	11.05	+ 0.733	11.58	+ 0.549	10.68	+ 0.476	11.10	+ 0.602
30 months	10.27	+ 0.227	10.67	+ 0.547	10.15	+ 0.134	10.36	+ 0.299
36 months	10.79	+ 0.386	11.22	+ 0.700	10.50	+ 0.190	10.84	+ 0.426

Abbreviation: TAM, tamoxifen.
*Difference = tamoxifen minus placebo.

the beginning of the P-1 trial and is similar in pattern and number to the adherence data recently reported in a second large, multicenter chemoprevention trial of hormone replacement therapy for heart disease.[16] We have shown that there is only a small difference in the proportion of nonadherent participants in the tamoxifen and placebo groups and that the nonadherent women in both trial groups have generally similar key demographic, clinical, and HRQL variables. Given these considerations, it seems unlikely that a maximum difference of 3% in the HRQL follow-up rates between the two groups was sufficient to create a significant bias in our between-group comparisons.

HRQL adherence is closely related to treatment adherence. Based on the reasons for quitting treatment, it would seem that nonadherent women in both trial groups were those who were sensitive to the actual or possible occurrence of side effects caused by tamoxifen.

Much concern has been expressed about a potential relationship between tamoxifen use and the onset of depression.[17-21] Women who reported a history of depressive episodes or a history of treatment for nervous or mental disorders were not excluded from the trial. A brief eight-item affective screening questionnaire based on the CES-D and the Diagnostic Interview Schedule[22] was part of the baseline examination.[23] Using data from this brief screening instrument, local investigators were alerted to eligible participants showing signs of potentially serious affective distress at the baseline examination and caution was advised regarding their enrollment onto the trial. However, women who showed current signs of affective distress or depression were not routinely excluded from the trial.

With regard to the primary screening instrument used in the follow-up examinations, it has been pointed out that "the items in... (the CES-D) are generally related to affective distress but not to any particular psychiatric disorder."[11] For this reason, the numbers listed in Table 3 refer not to the prevalence of clinically diagnosable depressive disorders but, instead, to the prevalence of clinically significant affective distress that might be associated with a number of specific psychiatric disorders. However, if tamoxifen use was associated with the onset of clinically diagnosable depression, we would have expected to see a consistent excess of individuals scoring ≥ 16 on the CES-D in the tamoxifen group. No such consistent excess was observed. These findings agreed with the data from the mental health scale on the SF-36.

The MOS SF-36 served in this study as a measure of overall HRQL. For this initial report, we have presented data from the SF-36 in terms of two high-level component scores[12] and the eight basic subscales generally used in scoring this instrument.[9] Neither of these two methods of summarizing the SF-36 data demonstrated any clinically significant differences between the tamoxifen and placebo groups.

The first clear signs of consistent differences between the tamoxifen and placebo groups were observed in the SCL. In 19 out of 21 follow-up comparisons, the mean number of symptoms reported on the SCL were consistently different by age group (50 to 59 years > 35 to 49 years > 60+ years) and by trial group (tamoxifen > placebo). The absolute differences between the trial groups were relatively small and tended to be associated with the types of vasomotor, gynecologic, and sexual functioning symptoms previously reported for tamoxifen.[18,24,25]

The data from the MOS sexual functioning scale indicate that relatively small (< 4.0%) but consistent differences exist between the two groups in regard to the proportion of women reporting definite or serious problems in at least three specific domains of sexual functioning, sexual interest, arousal, and orgasm. These problems do not seem to be age group specific. Despite these findings for specific domains of functioning, there is no evidence that these problems result in a reduction of the overall proportion of women in the tamoxifen group who are sexually active.

Table 5. Symptoms Reported at Least Once Between Months 3 and 36 With the Largest Relative Difference Between Trial Arms

Age Group and Symptom	Placebo Arm Proportion (%)	Tamoxifen Arm Proportion (%)	RR (TAM/Placebo)
35-49 years			
Cold sweats	15.90	22.90	1.44
Vaginal discharge	46.29	62.55	1.35
Pain in intercourse	23.88	31.57	1.32
Night sweats	59.58	74.16	1.24
Hot flashes	65.54	81.28	1.24
50-59 years			
Cold sweats	16.11	27.00	1.68
Vaginal discharge	32.51	53.47	1.64
Genital itching	36.93	45.24	1.23
Night sweats	62.77	75.88	1.21
Bladder control (laugh)	47.67	56.94	1.19
≥ 60 years			
Vaginal bleeding	4.64	10.92	2.35
Vaginal discharge	19.82	45.81	2.31
Genital itching	32.05	40.96	1.28
Hot flashes	51.51	63.59	1.23
Bladder control (laugh)	49.88	56.49	1.13
Overall			
Vaginal discharge	34.13	54.77	1.60
Cold sweats	14.77	21.40	1.45
Genital itching	38.29	47.13	1.23
Night sweats	54.92	66.80	1.22
Hot flashes	65.04	77.66	1.19
Pain in intercourse	24.13	28.19	1.17
Bladder control (laugh)	46.65	52.51	1.13
Bladder control (other)	47.79	52.83	1.11
Weight loss	41.97	44.94	1.07
Vaginal bleeding	21.26	21.96	1.03

Abbreviation: TAM, tamoxifen.

Based on these data, we conclude that tamoxifen use is associated with an increase in specific vasomotor, gynecologic, and sexual functioning symptoms. At the same time, we did not observe any evidence that overall physical and emotional well being were significantly affected by these differences in the frequency of symptoms. We also found no evidence on the CES-D or the SF-36 mental health scale for an association in any age group between tamoxifen use and an increase in the proportion of women reporting clinically significant levels of affective distress and/or depression. How should clinicians integrate the results from the HRQL study data into decision-making and recommendations to women considering the use of tamoxifen in the setting of prevention? As demonstrated by the SCL data from the placebo group of the trial, many symptoms experienced by women who participated in this study are age and menopause related and exist independent of the use of tamoxifen. However, several symptoms are substantially more frequent in women using tamoxifen; these include vasomotor symptoms (cold sweats, night sweats, and hot flashes), vaginal discharge, and genital itching. Women need to be informed

Table 6. Proportion of Women Reporting Hot Flashes in Tamoxifen Arm and RR Compared to Placebo Arm by Age Group and Examination

	Age Group							
	35-49 Years		50-59 Years		≥ 60 Years		Overall	
Examination	TAM	RR*	TAM	RR*	TAM	RR*	TAM	RR*
Baseline	0.258	0.959	0.533	0.989	0.268	1.030	0.346	0.991
3 months	0.581	1.588	0.761	1.241	0.511	1.413	0.616	1.399
6 months	0.610	1.666	0.765	1.268	0.503	1.481	0.626	1.455
12 months	0.614	1.525	0.740	1.273	0.460	1.412	0.606	1.396
18 months	0.613	1.510	0.715	1.239	0.419	1.461	0.586	1.387
24 months	0.622	1.457	0.681	1.199	0.388	1.311	0.570	1.322
30 months	0.627	1.362	0.642	1.206	0.330	1.177	0.541	1.265
36 months	0.627	1.414	0.667	1.276	0.364	1.362	0.560	1.348

Abbreviation: TAM, tamoxifen.
*RR = TAM/placebo.

of these possible symptoms. Weight gain and depression, two clinical problems anecdotally associated with tamoxifen treatment in women with breast cancer, did not increase in frequency in this large placebo-controlled trial of healthy women. This is good news that must also be communicated to women. An informed discussion with a woman considering tamoxifen therapy should include these points in the risk/benefit discussion.

Disclosure of likely and unlikely symptoms should prepare a woman for what she might experience and reduce her anxiety or concerns should she begin preventive therapy. Without the detailed evaluation of HRQL data obtained in the P-1 trial, we would not be able to provide this level of information and reassurance to women considering preventive therapy. In addition, the setting of preventive therapy differs considerably from the treatment of breast cancer. Therefore, if a woman experiences untoward symptoms after starting tamoxifen treatment, the medication can be discontinued if the symptoms cannot be controlled or her personal assessment of the risks and benefits changes.

Table 7. Proportion of Women Reporting Vaginal Discharge in Tamoxifen Arm and RR Compared to Placebo Arm by Age Group and Examination

	Age Group							
	35-49 Years		50-59 Years		≥ 60 Years		Overall	
Examination	TAM	RR*	TAM	RR*	TAM	RR*	TAM	RR*
Baseline	0.201	0.957	0.135	1.041	0.058	0.907	0.138	0.975
3 months	0.379	1.549	0.308	2.023	0.275	3.665	0.326	1.972
6 months	0.391	1.686	0.302	1.931	0.269	3.057	0.327	1.973
12 months	0.380	1.700	0.304	1.973	0.262	3.333	0.321	2.020
18 months	0.363	1.558	0.278	2.251	0.252	3.029	0.303	1.961
24 months	0.341	1.797	0.272	1.991	0.238	2.994	0.288	2.052
30 months	0.325	1.633	0.282	2.404	0.246	3.075	0.288	2.083
36 months	0.316	1.671	0.264	2.332	0.241	3.096	0.277	2.095

Abbreviation: TAM, tamoxifen.
*RR = TAM/placebo.

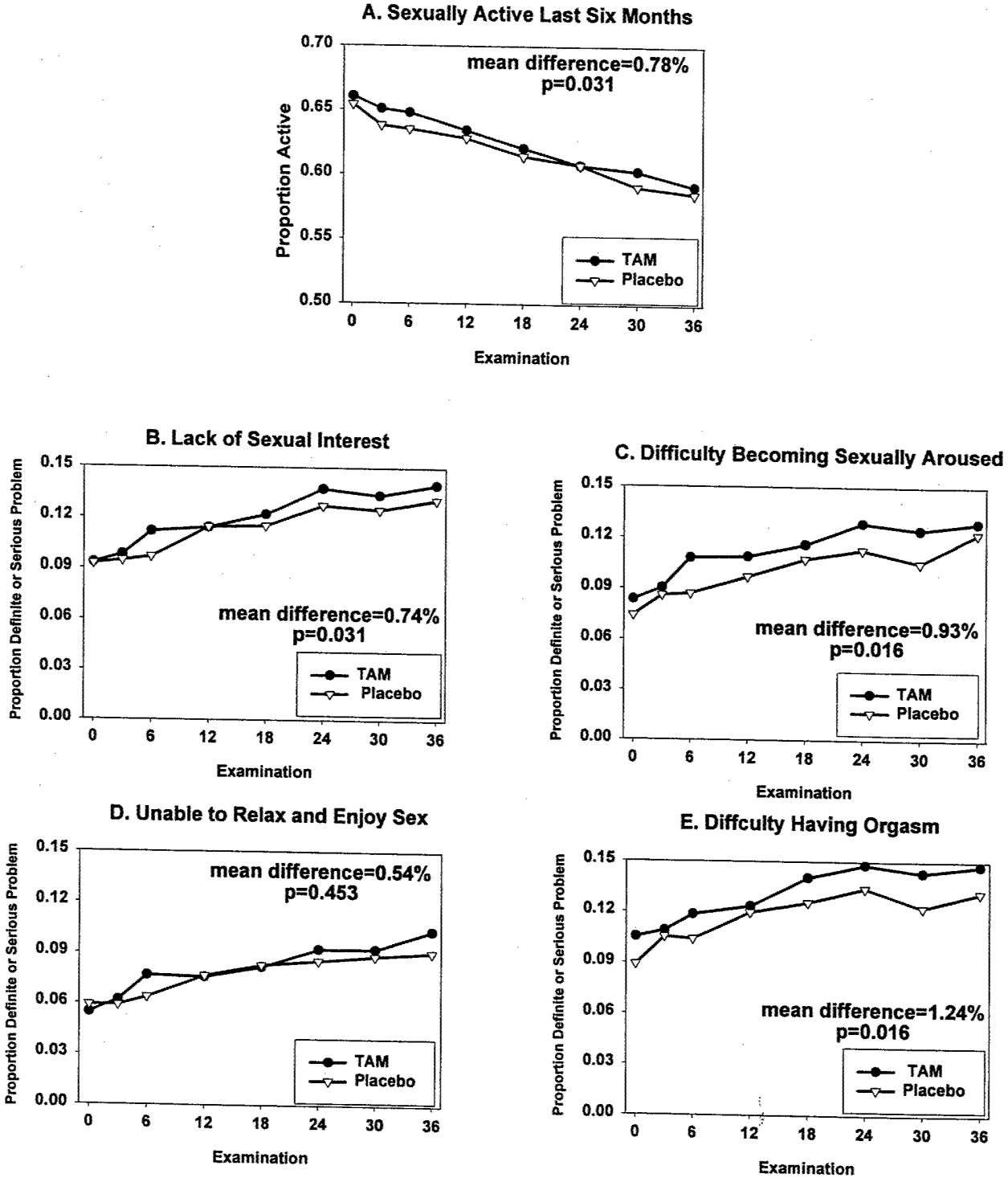


Fig 4. Proportion of women in the tamoxifen group and placebo group reporting a definite or serious problem in past 4 weeks on MOS sexual functioning scale (B through E, women who reported being sexually active in last 6 months).

The current report is a brief overview of the P-1 study HRQL data that focuses on important clinical and functional implications of tamoxifen use for women's overall HRQL. It will be supplemented in the future by a series of additional methodologic and clinical reports that will provide in-depth analyses of the data obtained from each one of the several P-1 study HRQL instruments.

ACKNOWLEDGMENT

We thank Carol Redmond, DSc, University of Pittsburgh; Leslie Ford, MD, National Cancer Institute, Bethesda, MD; Carol Moinpour, PhD, Southwest Oncology Group Statistical Center; John E. Ware, Jr, New England Medical Center, Boston, MA; David Cella, Northwestern University, Chicago, IL; Sheela Goshal and Wei Chen, NSABP Biostatistical Center; and members of the NSABP Prevention Quality of Life Committee.

REFERENCES

1. Fisher B, Costantino JP, Wickerham L, et al: Tamoxifen for the prevention of breast cancer: A report from the NSABP P-1 study. J Natl Cancer Inst 90:1371-1388, 1998

2. Fisher B, Costantino J: Highlights of the NSABP Breast Cancer Prevention Trial. Cancer Control 4:78-86, 1997

3. Ganz PA, Day R, Ware JE, et al: Base-line quality-of-life assessment in the National Surgical Adjuvant Breast and Bowel Project Breast Cancer Prevention Trial. J Natl Cancer Inst 87:1372-1382, 1995

4. Fisher B, Redmond C: Fraud in breast cancer trials. N Engl J Med 330:1458-1460, 1994

5. Fisher B, Anderson S, Redmond C, et al: Reanalysis and results after 12 years of follow-up in a randomized clinical trial comparing total mastectomy with lumpectomy with or without irradiation in the treatment of breast cancer. N Engl J Med 333:1456-1461, 1995

6. Ganz PA, Day R, Costantino JP: Compliance with quality of life data collection in the NSABP breast cancer prevention trial. Stat Med 17:613-622, 1998

7. Daniel WW: Applied Non-Parametric Statistics. Boston, MA, PWS-Kent Publishing Co, 1990

8. Deshpande JV, Gore AP, Shanubhougue A: Statistical Analysis of Non-Normal Data. New York, NY, John Wiley & Sons, 1995

9. International Resource Center for Health Care Assessment: How to Score the SF-36 Health Status Survey. Boston, MA, New England Medical Center, 1991

10. Radloff LS: The CES-D scale: A self-report depression scale for research in the general population. Appl Psychol Meas 1:385-401, 1977

11. Roberts RE, Vernon SW: The Center for Epidemiologic Studies Depression Scale: Its use in a community sample. Am J Psychiatry 140:41-46, 1983

12. Ware JE, Kosinski M, Keller SD: SF-36 Physical and Mental Summary Scales: A User's Manual (3rd Printing Revised). Boston, MA, The Health Institute, New England Medical Center, 1994

13. Fisher B: A commentary on endometrial cancer deaths in tamoxifen-treated breast cancer patients. J Clin Oncol 14:1027-1039, 1996

14. Fisher B, Costantino JP, Redmond CK, et al: Endometrial cancer in tamoxifen-treated breast cancer patients: Findings from NSABP B-14. J Natl Cancer Inst 86:527-537, 1994

15. Gorin MB, Day R, Costantino JP, et al: Long term tamoxifen citrate and potential ocular toxicity. Am J Ophthalmol 125:493-501, 1998

16. Hulley S, Grady D, Bush T, et al: Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. JAMA 280:605-613, 1998

17. Cathacart CK, Jones SE, Pumroy CS, et al: Clinical recognition and management of depression in node negative breast cancer patients treated with tamoxifen. Breast Cancer Res Treat 27:277-281, 1993

18. Love RL, Cameron L, Connell BL, et al: Symptoms associated with tamoxifen treatment in postmenopausal women. Arch Intern Med 151:1842-1847, 1991

19. Shariff S, Cumming CE, Lees A, et al: Mood disorder in women with early breast cancer taking tamoxifen, an estradiol receptor antagonist: An unexpected effect? Ann N Y Acad Sci 761:365-368, 1995

20. Moredo Anelli T, Anelli A, Tran KN, et al: Tamoxifen administration is associated with a high rate of treatment-limiting symptoms in male breast cancer patients. Cancer 74:74-77, 1994

21. Pluss JL, Dibella NJ: Reversible central nervous system dysfunction due to tamoxifen in a patient with breast cancer. Ann Intern Med 101:652, 1984

22. Robins LN, Helzer JE, Croughan J, et al: National Institute of Health Diagnostic Interview Schedule: Its history, characteristics and validity. Arch Gen Psychiatry 35:837-846, 1978

23. Burnam MA, Wells KB, Leake B, et al: Development of a brief screening instrument for detecting depressive disorders. Med Care 26:775-789, 1988

24. Fisher B, Dignam J, Bryant J, et al: Five versus more than five years of tamoxifen therapy for breast cancer patients with negative lymph nodes and estrogen receptor-positive tumors. J Natl Cancer Inst 88:1529-1542, 1996

25. Fisher B, Costantino J, Redmond C, et al: A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor-positive tumors. N Engl J Med 320:479-484, 1989

ARTICLES

Tamoxifen and Depression: More Evidence From the National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention (P-1) Randomized Study

Richard Day, Patricia A. Ganz, Joseph P. Costantino

Background: **Concerns have been raised that tamoxifen may be associated with depression. To investigate this question, we examined the psychological effects of tamoxifen treatment for breast cancer prevention on women at different levels of risk for clinical depression who were enrolled in the National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention (P-1) Study.** *Methods:* **A total of 11 064 women were randomly assigned to receive for 5 years daily doses of 20 mg of tamoxifen or placebo in the P-1 study, a multicenter, double-blind, placebo-controlled chemoprevention trial. Each woman was prospectively assessed for depression risk on the basis of medical history items collected at the baseline examination and placed in a high-, medium-, or low-risk group. Every 6 months, for a total of 36 months, the participants were assessed for depressive symptoms by completing the Center for Epidemiological Studies—Depression (CES-D) questionnaire. Scores of 16 or higher were indicative of an episode of affective distress. Differences between the risk groups and treatment arms were analyzed by logistic regression. All statistical tests were two-sided.** *Results:* **Women in the higher risk depression groups were more likely to score 16 or higher on the CES-D (percent follow-up examinations with a score of ≥16: high-risk group = 35.7%, with 95% confidence interval [CI] = 32.5% to 38.9%; medium-risk group = 19.2%, with 95% CI = 18.1% to 20.3%; and low-risk group = 8.7%, with 95% CI = 8.3 to 9.1%) and to have these scores more frequently and for longer periods than women in the lower risk groups. Within each depression risk group, there was no difference in the proportion of women scoring 16 or higher by treatment assignment (tamoxifen versus placebo) (odds ratio = 0.98; 95% CI = 0.93 to 1.02). A *post-hoc* analysis indicated that the lack of a tamoxifen effect was not a result of differential missing data.** *Conclusions:* **Physicians need not be overly concerned that treatment with tamoxifen will increase the risk for or exacerbate existing depression in women. Nevertheless, physicians should continue to screen for and treat or refer potential cases of depression encountered in routine clinical practice. [J Natl Cancer Inst 2001;93:1615–23]**

Concern regarding an association between clinical depression and tamoxifen, when used as an adjuvant treatment or preventative agent for breast cancer, has been voiced by a number of investigators *(1–5)* and continues to be discussed in regulatory agencies, such as the U.S. Food and Drug Administration. Furthermore, the *Physician's Desk Reference (6)* lists "depression" as an infrequent adverse reaction to tamoxifen. Although previous studies *(1–5)* used breast cancer patients to address tamoxifen use and depression, the studies had a number of weaknesses, including the lack of a clear definition of depression and a failure to control for the potential confounding effects of illness diagnosis, the side effects of chemotherapy (e.g., premature menopause), or normal aging. Previously, two double-blind, placebo-controlled studies of the effects of tamoxifen in postmenopausal women *(7,8)* found no association of tamoxifen with depression. We believe that some of the concern over the relationship between tamoxifen and depression arises from the idea that, because hormone replacement therapy has positive effects on mood and tamoxifen has antiestrogenic activity *(9–11)*, tamoxifen, therefore, has negative effects on mood.

The completion of the Breast Cancer Prevention (P-1) Study of the National Surgical Adjuvant Breast and Bowel Project (NSABP) provides an opportunity to investigate the association between tamoxifen and depression in greater detail. The P-1 study was a multicenter, double-blind, placebo-controlled chemoprevention trial. The primary objective of the study was to evaluate whether 5 years of tamoxifen therapy would reduce the incidence of invasive breast cancer in women at an increased risk for the disease. The secondary objectives of the study included the assessment of the incidence of ischemic heart disease, bone fractures, and other negative health events, such as depression, that might be associated with tamoxifen therapy. Eligible participants were randomly assigned to receive 20 mg daily of tamoxifen or a placebo for 5 years. Detailed reports on the rationale, planning, design, and clinical outcome of the P-1 study are available elsewhere *(12–16)*.

In our initial publication on the health-related quality of life (HRQL) *(16)* of all subjects in the P-1 study, we did not find a difference between the treatment groups (tamoxifen versus placebo) on the Center for Epidemiological Studies—Depression (CES-D) Scale *(17)* or the SF-36 Mental Health Scale *(18)*. It is known, however, that vulnerability to clinically identifiable forms of depression is not uniformly distributed in the general female population but, instead, clusters in high-risk groups of women *(19)*. This vulnerability to depression may be inherited, suggesting a genetic or familial origin, or it may be related to certain psychological predispositions, such as a low self-esteem, a poor resistance to stress, or a pessimistic view of the world. We

Affiliations of authors: R. Day, J. P. Costantino, National Surgical Adjuvant Breast and Bowel Project Biostatistical Center, Pittsburgh, PA; P. A. Ganz, Jonsson Comprehensive Cancer Center, Los Angeles, CA.

Correspondence to: Richard Day, Ph.D., Department of Biostatistics, Graduate School of Public Health, 130 DeSoto St., University of Pittsburgh, Pittsburgh, PA 15261 (e-mail: day@nsabp.pitt.edu).

See "Notes" following "References."

were concerned that the potential negative effects of tamoxifen for women at high risk for depression may have been masked in our previous analysis *(16)* because of the simultaneous inclusion of a larger group of less vulnerable (i.e., low-risk) participants.

In this study, we investigated the effects of tamoxifen on women at different levels of risk for depression. Specifically, we were interested in whether tamoxifen treatment was associated with the onset or prolongs the length of existing episodes of clinically diagnosable depression in women at high risk for depression.

Patients and Methods

Participant Cohort and HRQL Data

This article covers the baseline and first 36 months of follow-up data (collected at 6-month intervals) on the same 11 064 women used in the initial HRQL report *(16)* from the P-1 study. The P-1 participants ranged in age from 35 years to 79 years (mean ± standard deviation = 53.8 ± 9.2 years), were predominantly white (95.6%), were well educated (≥ some college = 64.9%), and were currently employed (full- or part-time = 64.7%) in a professional or technical field (67.9%). A detailed description of this cohort of participants and the P-1 HRQL instruments was reported previously *(14,16)*. All investigations conducted in the P-1 study were approved by review boards at each institution and were in accord with an assurance filed with and approved by the U.S. Department of Health and Human Services *(12)*. All of the participants provided written informed consent.

Defining Depression

Depressive disorders, as defined by the current psychiatric nomenclature in the *Diagnostic and Statistical Manual of Mental Disorders: DSM- IV* (DSM-IV) *(20)*, come in a variety of forms that differ on the basis of the number, severity, and persistence of symptoms. The majority of clinically diagnosable episodes of depression involve one of three disorders—major depression, dysthymia, or bipolar disorder *(19)*. Major depression involves an illness episode lasting at least 2 weeks that includes mood disturbance (dysphoria) and at least four of the following symptoms: sleep disturbance, change in psychomotor activity, loss of ability to experience pleasure and interest, fatigue, feelings of worthlessness or guilt, difficulty in concentrating, and a preoccupation with death or a wish to die. These symptoms must be associated with a clear impairment in social functioning. Dysthymic disorder or dysthymia is a chronic illness lasting at least 2 years. Dysthymia does not show the same levels of social impairment found in major depression, but it does involve mood disturbance (dysphoria) and a loss of the ability to experience pleasure and interest in usual activities, together with some of the other symptoms used to define major depression. Individuals diagnosed with dysthymia often experience episodes of major depression during their lifetime. DSM-IV distinguishes bipolar disorders from depressive disorders. Bipolar disorders have dramatic clinical manifestations that involve one or more episodes of hypomania during an individual's lifetime alternating with illness episodes that fit the criteria for major depression disorder.

Depression was previously defined by the Research Diagnostic Criteria (RDC) *(21)*, a nonclinical forerunner of the current DSM-IV criteria. The RDC used similar criteria as the DSM-IV to define "major depression" but, unlike the DSM-IV, also included criteria to define "minor depression" (nonpsychotic episodes of illness characterized by a prominent and sustained dysphoria but lacking all of the symptomatic features of major depression). Although important historically, the RDC has been superseded by the DSM-IV.

One of the problems associated with the definition of depression is that, in addition to these diagnosable clinical entities, there are multiple sources of affective distress that may result in short-term or self-limiting expressions of depressive symptoms without meeting the DSM-IV criteria outlined above. The best available data on rates of clinically diagnosable depressive disorders in the U.S. general population come from the National Institute of Mental Health's Epidemiological Catchment Area (ECA) study *(19)*. ECA study investigators found that, even though clinically diagnosable depressive disorders are relatively rare, usually affecting only 5%–6% of the general female population during any 12-month period, the reporting of depressive symptoms is reasonably frequent, with 35.7% of the women in the ECA study *(19)* reporting having experienced a period of dysphoria (feeling sad or blue) lasting at least 2 weeks. These expressions of affective distress, which fail to meet the clinical criteria for major depression, dysthymia, or bipolar illness, are often associated with occurrences such as uncomplicated grief, medical illness and other life events, or chronic difficulties *(22)*. Depressive symptoms may also occur secondary to other psychiatric illnesses (i.e., anxiety disorders or phobias), chronic medical conditions, or substance abuse.

Monitoring Depressive Symptoms in the P-1 Study

The primary instrument used to monitor depressive symptoms in the P-1 study was the CES-D *(17)*. This self-administered questionnaire was designed to be a brief, first-stage screen rather than a clinical diagnostic instrument. The CES-D is composed of 20 items, each of which is scored on a scale of 0–3. Higher scores reflect increased expression of affective distress, and a score of 16 or higher is most often used as the cutoff point for likely cases of clinical depression *(17,23,24)*.

Two problems are associated with the use of the CES-D alone to screen for clinically diagnosable episodes of depression. First, questions on the CES-D inquire only about the past 7 days, collecting little information on the length of time that a symptom has been present. Second, the CES-D collects information only on symptoms and not the degree of social impairment experienced by the respondent. Consequently, scores above the CES-D clinical cutoff point of 16 tend to include a substantial proportion of distressed individuals—perhaps upwards of one half or more—who do not meet the clinical criteria for major depression, dysthymia, or bipolar illness *(24,25)*.

Estimating Depression Risk in P-1 Study Participants

The eligibility criteria for the P-1 study permitted, at the discretion of the local site investigator, the inclusion of women with evidence of clinical depression. Twenty to 22% of the participants scored 16 or higher on the CES-D at least once during any 12-month period of the P-1 study. This percentage exceeds the expected general population rates [5%–6% *(19)*] of clinically diagnosable depressive disorders over a 12-month period by 3.5–4.0 times, indicating that it is necessary to distinguish between clinically diagnosable episodes of depression and depressive symptoms that are secondary to other types of physical and psychiatric illnesses or a consequence of social conditions that produce short-term, self-limiting expressions of affective distress. The preferred means to make such a distinction would be a standardized psychiatric interview, such as the Schedule for Affective Disorders and Schizophrenia—Lifetime Version *(26)* or the Diagnostic Interview Schedule *(19)*. However, in the absence of such an interview, the best single indicator of risk for a future episode of major depression, dysthymia, or bipolar disorder in the P-1 study data is a medical history of treatment for these disorders *(27–30)*.

The ECA study *(19)* found that the mean age at onset for major depressive disorders in the general population was 27 years, with approximately 89% of all first depressive episodes occurring before age 35 years, which was the lower age limit of the participants in the P-1 study. Medical history information, collected on a one-time-only basis as a part of the baseline entry and eligibility assessment of all P-1 study participants, included three self-reported items regarding depression: 1) a medical history of depression, 2) current or previous prescriptions for antidepressant medications, and 3) extended periods (≥12 months) of dysphoric mood (i.e., "depressed or sad most days"). If a participant gave a positive answer to the medical history or the medication question, the interviewer obtained dates of treatment, physicians' names, specific modalities of treatment, and date of last medication dose to assess the consistency and appropriateness of the information provided.

These three medical history items were used in the current study to prospectively estimate each participant's risk of experiencing a clinically diagnosable episode of depression. A simple three-level risk score was determined for each P-1 study participant, depending on whether they endorsed 0 (low risk), 1 or 2 (medium risk), or 3 (high risk) of the medical history items regarding depression in the Entry/Eligibility Form. We hypothesized that women with higher scores on this simple depression risk scale would experience more severe and persistent episodes of affective distress and would be more likely to receive a clinical diagnosis of depression. Moreover, if tamoxifen was associated with the onset and/or prolonged the length of depressive episodes in the high-risk (i.e., more vulnerable) group, it should be apparent from longitudinal differences in the proportion of P-1 study participants in the treatment groups (tamoxifen versus placebo) who scored 16 or higher on the CES-D.

Statistical Analysis

CES-D scores were analyzed as above or below the clinical cutoff of 16 or higher. Binary logistic regression was the primary method of statistical analysis used in this study. Estimated odds ratios (ORs), confidence intervals (CIs), and *P* values are provided for all inferential analyses. Cox regression analysis was used to investigate the effects of treatment and depression risk on the time to the first CES-D with a score of 16 or higher, and Kaplan–Meier curves are provided for these data. When the CES-D data were handled as a continuous variable, nonparametic equivalents to a one-way analysis of variance (i.e., Kruskal–Wallis test) were used because it is unusual for CES-D scores to be normally distributed. Graphic presentations include 95% CIs on observed proportions to provide the reader with visual criteria for the magnitude of potential variation. Reported *P* values are all two-sided and have not been adjusted for multiple statistical comparisons. Instead, we have chosen to focus on consistent patterns of findings rather than on individual statistical tests in forming our conclusions. We also avoided the use of statistical methods for imputation of missing data points in the primary data because the data did not meet the strong assumptions that normally underlie such procedures (e.g., MCAR [i.e., Missing Completely at Random]/ MAR [i.e., Missing at Random]). Analyses were carried out with the use of Minitab (Version 13; Minitab, State College, PA) and Egret (Version 1.0; Cytel Corp., Cambridge, MA).

RESULTS

Depression Risk

To determine whether there was an association between depression and tamoxifen treatment in participants of the P-1 study, we first calculated the depression risk score from the frequency of responses to each one of the medical history items (Table 1). The three components of this score were only moderately intercorrelated. The highest correlation occurred between a history of illness and antidepressant medications ($r = .564$; $P<.001$), followed by history of illness and persistent dysphoric mood ($r = .369$; $P<.001$) and medications and dysphoric mood ($r = .269$; $P<.001$). Overall depression risk, measured by the data from this study, was not statistically significantly related to the participants' risk of breast cancer, as measured by the Gail risk model *(12,31)*.

The construct validity of this depression risk score was evaluated, in part, with the use of the social and demographic factors associated with clinically diagnosable depressive disorders in the ECA study *(19)*. Table 2 shows the distribution of the P-1 study participants according to the three-level depression risk scale on seven demographic variables, which approximate those associated with clinically diagnosable depression in the ECA study *(19)*. All of these variables, except education, showed a statistically significant dose–response relationship to the depression risk scores in terms of the direction and intensity of the association.

CES-D Data

Fig. 1, a, shows the proportion of the participants in each depression risk group who scored above the clinical cutoff of 16 or higher on the CES-D Scale at baseline and at each of the follow-up examinations. A consistent, positive dose–response relationship was seen between depression risk, as determined on the basis of the medical history items, and the proportion of participants scoring 16 or higher on the CES-D Scale at each scheduled examination. For each depression risk group, Table 3 shows the mean proportion of follow-up examinations with scores of 16 or higher and the distribution of the maximum and the overall scores on CES-D examinations above the clinical cutpoint. A positive dose–response relationship was also observed between depression risk group and proportion of respondents who scored 16 or higher on sequential CES-D examinations. In the high-risk depression group, for example, 21.2% of the respondents scored 16 or higher on three or more sequential CES-D examinations, compared with 9.7% for the medium-risk group and 3.5% for the low-risk group (data not shown). These findings confirm the expectation that participants in the higher depression risk groups (high>medium>low), on average, tend to experience more persistent and severe episodes of affective distress.

We next analyzed the CES-D data from each depression risk group by treatment group (tamoxifen versus placebo) (Fig. 1, b–d; Table 4). After adjustment for examination and risk group, the results of a logistic regression found that there was a statistically nonsignificant effect for the tamoxifen group compared with the placebo group (OR = 0.98; 95% CI = 0.93 to 1.02; $P = .32$). These analyses indicate that treatment group is not statistically associated with the proportion of women scoring above the CES-D clinical cutoff of 16 or higher in any of the three depression risk groups. Furthermore, after adjustment for depression risk group, an analysis of variance found that

Table 1. Distribution of self-reported risk factors for clinical depressive disorders at baseline examination among participants of the National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention (P-1) Study

Risk group (items endorsed)	History of depression	Antidepressant medications	Persistent dysphoria	Risk factor pattern*		0–3 risk factors†	
				No.	%	No.	%
Low (0)	No	No	No	7964	72.0	7964	72.0
Medium (1)	No	No	Yes	621	5.6		
	No	Yes	No	668	6.0	1628	14.7
	Yes	No	No	339	3.1		
Medium (2)	No	Yes	Yes	120	1.1		
	Yes	No	Yes	202	1.8	953	8.6
	Yes	Yes	No	631	5.7		
High (3)	Yes	Yes	Yes	519	4.7	519	4.7
Total				11 064	100.0	11 064	100.0

*Depression risk groups were assigned on the basis of the participants' response to three medical history questions: 1) history of depression, 2) use of antidepressant medication, and 3) persistent mood disturbance (dysphoria). Each positive answer was worth 1 point. Participants with a score of 0 were assigned to the low-risk group, those with a score of 1–2 to the medium-risk group, and those with a score of 3 to the high-risk group.

†Number and percent of participants endorsing 0, 1, 2, or 3 depression risk factors.

Table 2. Distribution of NSABP P-1 participants on ECA study social and demographic correlates of clinically diagnosed depressive disorders by depression risk score*

Sociodemographic item	Depression risk score†			Odds ratio‡	95% confidence interval on odds ratio
	Low, %	Medium, %	High, %		
Marital status: divorced/separated	11.1	17.7	23.5	1.63	1.50 to 1.98
Employment status: not working	4.4	7.9	12.2	1.78	1.58 to 2.01
Visited doctor within last 3 mo	71.0	76.4	84.4	1.39	1.28 to 1.51
Hospitalized within last 5 y	42.7	48.6	54.9	1.27	1.19 to 1.36
Age: ⩾60 y	29.9	27.4	24.1	0.87	0.81 to 0.94
Education: >high school	66.6	66.7	70.0	1.04	0.97 to 1.12
Income: >median	46.1	37.6	31.5	0.72	0.67 to 0.77

*NSABP P-1 = Natural Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention (P-1) Study; ECA = National Institutes of Mental Health's Epidemiological Catchment Area study *(19)*.

†Depression risk groups were assigned on the basis of the participants' responses to three medical history questions: 1) history of depression, 2) use of antidepressant medication, and 3) persistent mood disturbance (dysphoria). Each positive answer was worth 1 point. Participants with a score of 0 were assigned to the low-risk group, those with a score of 1–2 to the medium-risk group, and those with a score of 3 to the high-risk group.

‡Odds ratios were determined by binary logistic regression; $P<.001$ for all groups compared with referent groups, except for education, where $P=.235$.

there was no difference in the mean individual proportion of follow-up examinations above the clinical cutoff in each treatment arm.

The Kaplan–Meier plot in Fig. 2 shows the relationship between assigned treatment (placebo versus tamoxifen) and depression risk group (high, medium, or low) for the time from randomization until the first CES-D examination with a score exceeding the clinical cutoff of 16 or higher. The results of Cox proportional hazards regression analysis with these data were statistically significant for depression risk group (likelihood

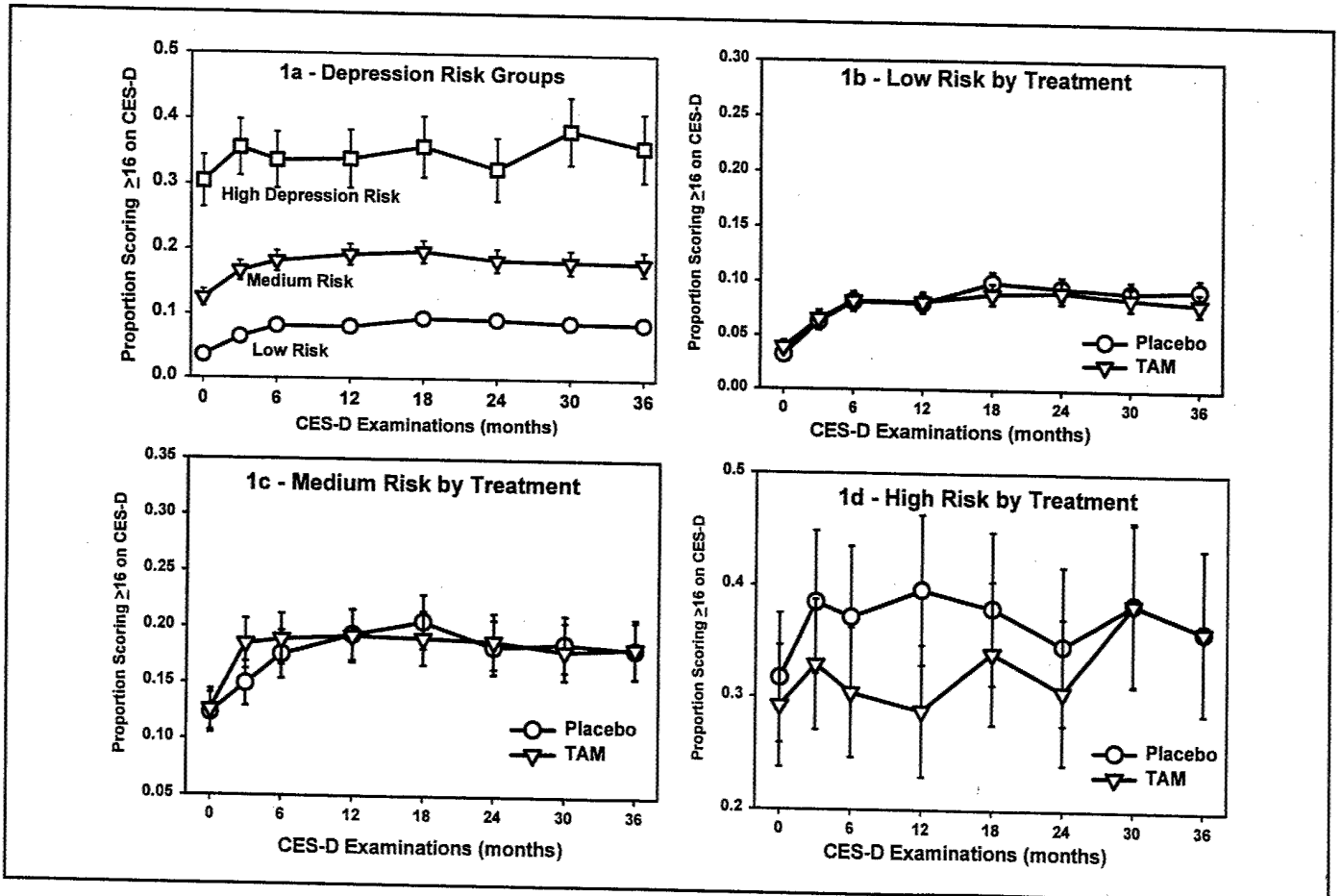


Fig. 1. Proportion of participants in the National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention (P-1) Study scoring 16 or higher on the Center for Epidemiological Studies—Depression (CES-D) Scale with 95% confidence intervals by depression risk groups (low, medium, or high) (**a**) and by depression risk group and treatment assignment (placebo versus tamoxifen [TAM]) (**b–d**). Depression risk groups were assigned on the basis of the participants' responses to three medical history questions: 1) history of depression, 2) use of antidepressant medication, and 3) persistent mood disturbance (dysphoria). Each positive answer was worth 1 point. Participants with a score of 0 were assigned to the low-risk group, those with a score of 1–2 to the medium-risk group, and those with a score of 3 to the high-risk group.

Table 3. Distribution of Center for Epidemiological Studies—Depression (CES-D) Scale variables for NSABP P-1 participants who scored above the clinical cutoff of 16 or higher by depression risk group*

CES-D variable	Depression risk group†		
	Low	Medium	High
% follow-up examinations in which participants scored ≥16‡			
Mean	0.087	0.192	0.357
95% CI for mean	0.083 to 0.091	0.181 to 0.203	0.325 to 0.389
Maximum score ≥16‡			
Median	22	24	27
Mean	23.97	25.61	28.58
95% CI for mean	23.66 to 24.28	25.16 to 26.06	27.62 to 29.54
All scores ≥16‡			
Median	20	21	22
Mean	21.52	22.49	23.74
95% CI for mean	21.30 to 21.74	22.17 to 22.81	23.10 to 24.38

*The CES-D is a self-administered questionnaire, composed of 20 items, each of which is scored on a scale of 0–3. Higher scores reflect increased expression of affective distress, and a total score of 16 or higher is used as the cutoff point for likely cases of clinical depression *(17,23,24)*. NSABP P-1 = National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention (P-1) Study; CI = confidence interval.

†Depression risk groups were assigned on the basis of the participants' responses to three medical history questions: 1) history of depression, 2) use of antidepressant medication, and 3) persistent mood disturbance (dysphoria). Each positive answer was worth 1 point. Participants with a score of 0 were assigned to the low-risk group, those with a score of 1–2 to the medium-risk group, and those with a score of 3 to the high-risk group.

‡There is a statistically significant difference between all groups (Kruskal–Wallis and analysis of variance: $P<.001$). "Maximum score ≥16" represents the highest single CES-D score ≥16 reported for an individual, whereas "All scores ≥16" summarizes all of the CES-D scores ≥16 reported for an individual.

ratio statistic [LRS] $P<.001$; hazard ratio [HR] = 1.88; 95% CI = 1.74 to 2.05), but they were statistically nonsignificant for both treatment arm effects (LRS $P = .988$; HR = 1.00; 95% CI = 0.92 to 1.09) and interaction effects (LRS $P = .575$; HR = 1.03; 95% CI = 0.92 to 1.16). The proportional hazards assumption for this analysis was confirmed.

Missing Data

We next assessed the association between missing data and depression risk group or sequential CES-D examination (Fig. 3, a). Logistic regression analysis based on the data in Fig. 3, a, indicated that depression risk group (OR = 1.17; 95% CI = 1.13 to 1.21; $P<.001$) and sequential examination (OR = 1.45; 95% CI = 1.44 to 1.46; $P<.001$) were both statistically significantly associated with missing CES-D data. Panels b–d in Fig. 3 show the proportion of participants completing the CES-D by depression risk and treatment groups. Logistic regression analysis by depression risk, controlling for sequential examination, indicates that, compared with placebo treatment, tamoxifen treatment was associated with higher proportions of missing data in the low-risk group (OR = 1.11; 95% CI = 1.06 to 1.16; $P<.001$) and the medium-risk group (OR = 1.12; 95% CI = 1.04 to 1.21; $P<.001$) but not in the high-risk group (OR = 0.99; 95% CI = 0.84 to 1.16; $P = .91$). If tamoxifen-associated depression were the primary cause of these missing data, we would have predicted a positive dose–response increase in the magnitude of the ORs from the lowest to the highest depression risk group.

We noted in our previous report *(15)* that it was difficult to continue to collect quality-of-life data after a participant had gone off treatment. However, participants in the P-1 study were asked about their primary reason for going off treatment, and their responses were recorded on an Off Therapy Form (OTF) that included "depression" as one of 10 specific response categories.

Of the 11 064 participants in this cohort, we collected an OTF for 3539 (80.8%) of 4382 women who missed at least one CES-D examination. The presence of an OTF showed a moderate positive correlation with the total number of missing CES-D examinations ($r = .62$; $P<.001$). The women who completed an OTF accounted for 12 693 (89.7%) of 14 149 missing

Table 4. Comparison (binary logistic regression) of the proportion of NSABP P-1 participants in each treatment group (tamoxifen versus placebo) who scored 16 or higher on the Center for Epidemiological Studies—Depression (CES-D) Scale by depression risk group and sequential examination*

Depression risk group†	Sequential examination							
	Baseline	3 mo	6 mo	12 mo	18 mo	24 mo	30 mo	36 mo
Low								
OR‡	1.22	1.04	1.01	1.02	0.88	0.96	0.93	0.86
95% CI	0.96 to 1.55	0.86 to 1.25	0.85 to 1.19	0.86 to 1.02	0.75 to 1.04	0.80 to 1.13	0.78 to 1.12	0.71 to 1.03
P	.10	.68	.91	0.86	.14	.60	.44	.11
Medium								
OR‡	1.03	1.29	1.10	0.99	0.91	1.04	0.96	1.01
95% CI	0.81 to 1.30	1.04 to 1.60	0.89 to 1.35	0.81 to 1.22	0.74 to 1.13	0.82 to 1.30	0.75 to 1.22	0.79 to 1.29
P	.84	.02	.39	.95	.40	.76	.72	.94
High								
OR‡	0.89	0.78	0.74	0.62	0.84	0.83	1.00	1.00
95% CI	0.61 to 1.30	0.54 to 1.14	0.50 to 1.09	0.41 to 0.92	0.56 to 1.26	0.54 to 1.28	0.65 to 1.54	0.64 to 1.57
P	.54	.21	.13	.02	.40	.40	.99	.99

*The CES-D is a self-administered questionnaire, composed of 20 items, each of which is scored on a scale of 0–3. Higher scores reflect increased expression of affective distress, and a total score of 16 or higher is used as the cutoff point for likely cases of clinical depression *(17,23,24)*. NSABP P-1 = National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention (P-1) Study; OR = odds ratio; CI = confidence interval.

†Depression risk groups were assigned on the basis of the participants' responses to three medical history questions: 1) history of depression, 2) use of antidepressant medication, and 3) persistent mood disturbance (dysphoria). Each positive answer was worth 1 point. Participants with a score of 0 were assigned to the low-risk group, those with a score of 1–2 to the medium-risk group, and those with a score of 3 to the high-risk group.

‡OR >1.0 indicates a greater proportion of women in the tamoxifen group.

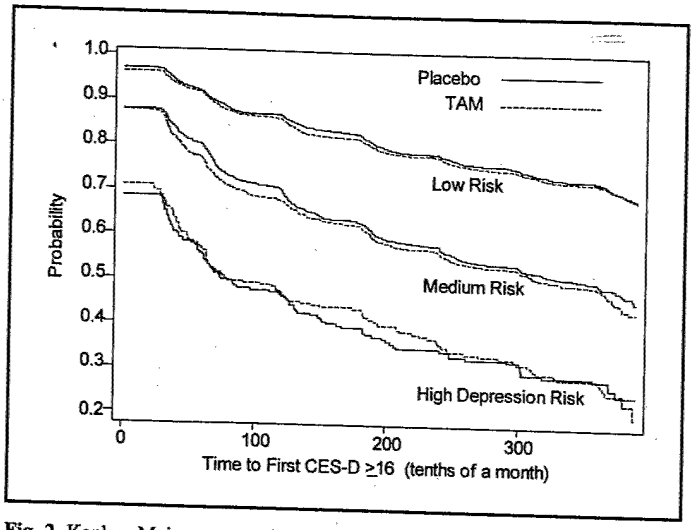


Fig. 2. Kaplan–Meier curves of time from randomization to first score of 16 or higher on the Center for Epidemiological Studies—Depression (CES-D) Scale by depression risk group (low, medium, or high) and treatment assignment (placebo versus tamoxifen [TAM]). Depression risk groups were assigned on the basis of the participants' responses to three medical history questions: 1) history of depression, 2) use of antidepressant medication, and 3) persistent mood disturbance (dysphoria). Each positive answer was worth 1 point. Participants with a score of 0 were assigned to the low-risk group, those with a score of 1–2 to the medium-risk group, and those with a score of 3 to the high-risk group. At 10 months, for the patients who received tamoxifen, in the low-risk group there were 3159 patients at risk of depression (proportion remaining = 0.864; 95% confidence interval [CI] = 0.853 to 0.875); in the medium-risk group there were 799 patients at risk (proportion remaining = 0.685; 95% CI = 0.659 to 0.711); and in the high-risk group there were 123 patients at risk (proportion remaining = 0.488; 95% CI = 0.427 to 0.549). At 30 months, for the patients who received tamoxifen, in the low-risk group there were 2233 patients at risk for depression (proportion remaining = 0.746; 95% CI = 0.732 to 0.760); in the medium-risk group there were 496 patients at risk (proportion remaining = 0.528; 95% CI = 0.499 to 0.557); and in the high-risk group there were 61 patients at risk (proportion remaining = 0.317; 95% CI = 0.258 to 0.376). At 10 months, for the patients who received the placebo, in the low-risk group there were 3190 patients at risk for depression (proportion remaining = 0.870; 95% CI = 0.859 to 0.881); in the medium-risk group there were 863 patients at risk (proportion remaining = 0.713; 95% CI = 0.688 to 0.738); and in the high-risk group there were 108 patients at risk (proportion remaining = 0.475; 95% CI = 0.412 to 0.538). At 30 months, for the patients who received the placebo, in the low-risk group there were 2326 patients at risk for depression (proportion remaining = 0.753; 95% CI = 0.738 to 0.767); in the medium risk group there were 544 patients at risk (proportion remaining = 0.535; 95% CI = 0.506 to 0.563); and in the high-risk group there were 59 patients at risk (proportion remaining = 0.316; 95% CI = 0.254 to 0.377).

CES-D examinations. Only 110 (3.1%) of these 3539 women reported that depression was the primary reason for their going off therapy. The most frequent reasons for going off therapy were nonmedical in nature (1667 women [47.1%]), perceived toxic effects (921 women [26.0%]), and various protocol and nonprotocol medical conditions (841 women [23.8%]).

Table 5 shows the distribution of women who reported that depression was their primary reason for going off treatment by treatment group and depression risk group. An analysis of these data using binary logistic regression found a statistically significant effect for depression risk group (OR = 2.37; 95% CI = 1.83 to 3.07; $P<.001$) and a statistically nonsignificant effect for treatment group (OR = 1.10; 95% CI = 0.75 to 1.62; $P = .63$), indicating that the cases of depression that lead women to quit their assigned treatment did not occur with a greater frequency in those in the tamoxifen arm.

Discussion

Tamoxifen is the most widely prescribed anticancer agent currently in use. It has been proven to be effective against breast cancer as an adjuvant treatment and in a preventative setting *(12,32)*. Given the widespread use of tamoxifen, it is important to fully investigate all of the potential side effects that may be associated with its administration, so that women, together with their physicians, can make an informed decision regarding its potential costs and benefits and its appropriateness for their individual situations.

This study is an extension of our earlier report *(16)* on the HRQL data from the NSABP P-1 study. Previously, we found no evidence for an association between tamoxifen treatment and depression in the overall P-1 study cohort. In this study, we recognized that vulnerability to clinically identifiable depressive disorders is not randomly distributed in the general female population and that the effects of tamoxifen on susceptible women in the P-1 study may have previously gone undetected.

Our initial problem was the *a priori* identification of subgroups of women with a potential clinical susceptibility for depression. Because the self-administered depression-screening form (CES-D) used in the P-1 study provides information on short-term symptoms of affective distress and is not intended for use as a diagnostic instrument *(17)*, we incorporated the participants' self-reported medical history of depression, use of prescription antidepressant medications, and experience of extended periods (>12 months) of dysphoric mood to assign clinical risk. On the basis of these data, women were prospectively assigned to one of three depression risk groups. We hypothesized that the higher a woman's depression risk group, the greater the likelihood that she would experience a clinically diagnosable episode of depression.

The P-1 study staff were trained to check the consistency and appropriateness of the self-reported data about prior treatment for depression and the use of antidepressant medications as a routine part of the medical screening procedure carried out during entry/eligibility interview. These procedures were designed specifically to minimize false-positive classification errors. However, there was little that the interviewer could do to detect false-negative classification errors in which a potential participant did not, for whatever reason, report the requested screening information. The overall effect of this inability to control for false-negative classification errors for the current study was to create a potential misclassification bias in which women at increased risk for depression may have been placed, at an unknown rate, in one of the lower risk groups. Although less than ideal, the effect of this bias is conservative in nature, operating to maintain the comparative validity of the most important high-risk depression group.

We found a statistically significant dose–response relationship between the level of the depression risk group (high> medium>low) and the proportion of the women in each depression risk group who scored above the clinical cutoff of 16 or higher on the CES-D at baseline and at every follow-up interview. In addition, women in the higher risk groups (high> medium>low) scored above the clinical cutoff on a greater proportion of their follow-up interviews and, on average, had higher maximum CES-D scores. Together, these data suggest that there was a dose–response effect, in which women in the higher depression risk groups (high>medium>low) were more likely to

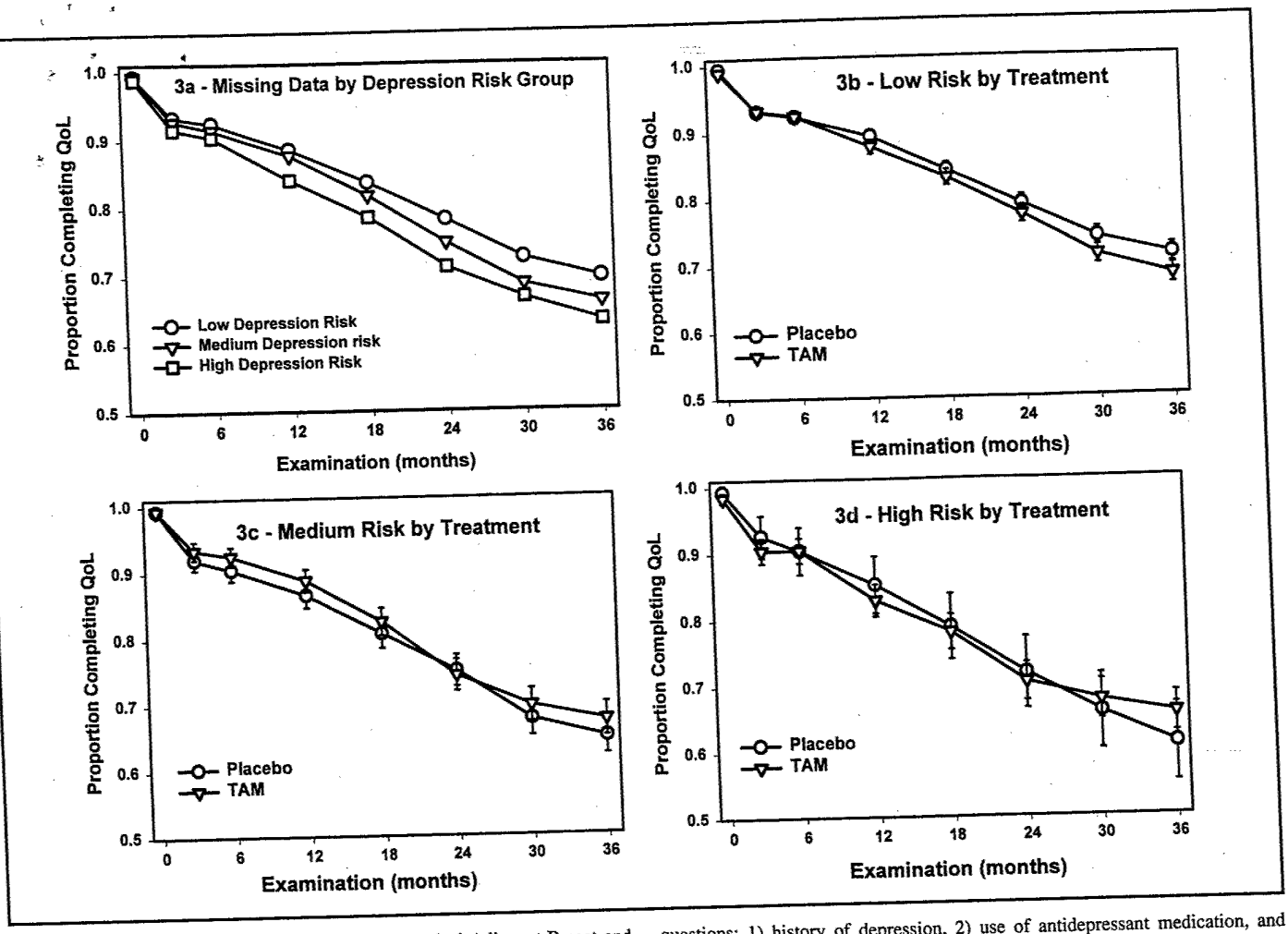


Fig. 3. Proportion of participants in the National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention (P-1) Study completing the health-related quality-of-life questionnaire by depression risk groups (low, medium, or high) (**a**) and by depression risk group and treatment assignment (placebo versus tamoxifen [TAM]) with 95% confidence intervals (**b–d**). Depression risk groups were assigned on the basis of the participants' responses to three medical history questions: 1) history of depression, 2) use of antidepressant medication, and 3) persistent mood disturbance (dysphoria). Each positive answer was worth 1 point. Participants with a score of 0 were assigned to the low-risk group, those with a score of 1–2 to the medium-risk group, and those with a score of 3 to the high-risk group.

Table 5. Reasons cited for going off treatment by depression risk* and treatment group

	Low risk		Medium risk		High risk		
Reasons cited for going off treatment	Placebo	Tamoxifen	Placebo	Tamoxifen	Placebo	Tamoxifen	Overall
Depression (No. of participants)	20	27	21	24	9	9	110
Other reasons (No. of participants)	1130	1275	416	431	83	94	3429
Depression as % of all off-treatment reasons	1.7	2.1	4.8	5.3	9.8	8.7	3.1

*Depression risk groups were assigned on the basis of the participants' responses to three medical history questions: 1) history of depression, 2) use of antidepressant medication, and 3) persistent mood disturbance (dysphoria). Each positive answer was worth 1 point. Participants with a score of 0 were assigned to the low-risk group, those with a score of 1–2 to the medium-risk group, and those with a score of 3 to the high-risk group.

experience clinically significant episodes of affective distress and that these episodes, on average, were more persistent and more severe than the episodes in the lower risk groups. Finally, we found that the distribution of social and demographic correlates (i.e., age, marital and employment status, educational level, and use of medical services) across the three depression risk groups defined in this study followed the same general patterns of risk previously identified in the ECA study of depression among the general population *(21)*. All of the above findings serve to support the validity of the risk assignments used in our study.

The primary test of our research question involved stratifying each depression risk group by treatment assignment (tamoxifen versus placebo) and comparing the corresponding proportions of women at each follow-up interview who scored above the clinical cutoff of 16 or higher on the CES-D. We found no effect of tamoxifen for any of the three depression risk groups.

Besides the lack of a positive association between tamoxifen use and depression, there are at least two possible alternative explanations for our negative findings: lack of statistical power and missing data. We carried out a *post-hoc* effect size analysis to determine the size of the difference between the treatment

arms that might have been detected. For our highest risk depression group (n = 519), we had an 80% chance of detecting at least a 37% (OR ≥1.37) increase between the two study arms in the proportions of women scoring above the CES-D clinical cutoff of 16 or higher at any single examination point. When a repeated measures design was used, we had sufficient power to detect a mean increase of 24% (OR ≥1.24) in the proportion of women in either arm scoring above the CES-D clinical cutoff *(33,34)*. We considered these to be acceptable levels of statistical power for the identification of clinically significant treatment effects in our high-risk depression group. The detectable ORs were, of course, even smaller for the low- and medium-risk depression groups.

We also assessed the contribution of missing data to explain the negative association between tamoxifen and depression in the P-1 study. An initial analysis showed that assigned depression risk was statistically significantly associated with missing data rates over the course of the study. If a tamoxifen-associated depression was the primary cause of these rates, we would have predicted that the tamoxifen treatment group in the higher depression risk groups would show a progressively greater differential off-treatment rate than the placebo group. This expectation was not confirmed by our data for the high-risk depression group.

In addition, we also examined the reasons given for going off the assigned treatment. There was a strong statistical association in the P-1 study between stopping assigned treatment and missing HRQL data *(16)*. An analysis of the reasons for going off treatment in 81% of the women with missing HRQL data resulted in the following observations: (*a*) Depression was cited as a relatively infrequent reason for going off treatment; (*b*) the higher the depression risk group, the greater the likelihood that depression was cited as the reason for going off treatment; and (*c*) within each depression risk group, depression was cited as the reason for going off treatment by similar proportions of women, regardless of treatment assignment. A separate report *(35)* has implemented a sensitivity analysis on these data with equally negative results. The findings in our report together with this sensitivity analysis indicates that there are no clear patterns in the missing data that serve to undermine the conclusions drawn from our primary analysis.

The results of our analysis strengthen our previous conclusion regarding lack of evidence for an association between tamoxifen use and depression in the P-1 study data by provisionally extending our findings to subgroups of women at a high risk for clinically identifiable episodes of depression. Clinically, these findings have two major implications. First, the evidence from NSABP's P-1 study does not lend support to the idea that tamoxifen should be considered to be a causal risk factor for the onset of depressive symptoms and/or the prolongation of depressive episodes that occur among treated women. Second, the findings of this study suggest that physicians need not automatically disqualify women as candidates for tamoxifen treatment simply because they report a history of depressive symptoms or prior treatment for a depressive disorder. Nevertheless, it is still essential that physicians carefully screen for affective disorders and treat or refer potential cases of depression encountered in routine clinical practice.

Finally, there are two important limitations on these conclusions that require discussion, one statistical and the other methodological. Statistically, it was the large size of the P-1 study that permitted us to identify and carry out stratified analyses of groups of women with a differential risk for depression. However, we also noted that there were limits on our statistical power to detect an increase in the proportion of women reporting clinically significant levels of depressive symptoms on the CES-D, particularly in the high-risk depression group. For this reason, we cannot absolutely exclude the possibility that there may be rare cases in which women react negatively to tamoxifen treatment with potentially life-threatening depressions. Here, it is useful to recall that data on neuro-mood toxic effects were collected for P-1 study participants and periodically reviewed as part of the routine safety-monitoring procedures. Over the full course of the P-1 study, there were a total of three women who committed suicide, one woman from the placebo-treated group and two women from the tamoxifen-treated group, and there were no statistically significant differences in the distribution of women reporting suicidal ideation across the two trial arms.

The methodological limitations of this article (i.e., the lack of standardized psychiatric diagnoses and missing HRQL data) are primarily due to the fact that the goals of this study were secondary to the main clinical objectives that determined the design of the P-1 study. A more definitive analysis would require additional data from a potentially smaller, yet more focused study, in which an investigation of the relationship between clinical depression and tamoxifen treatment was the primary scientific objective. Such a study would have to have the following minimum features: (*a*) a double-blind, placebo-controlled, randomized design; (*b*) participants who are at high risk for breast cancer, rather than breast cancer patients (to avoid potential confounding due to clinical diagnosis and treatment); (*c*) participants who are stratified on a reliable measure of risk for affective disorder (e.g., lifetime diagnosis, Schedule for Affective Disorders and Schizophrenia—Lifetime Version); (*d*) periodic administration, in whole or in part, of a standardized psychiatric diagnostic instrument (e.g., Diagnostic Interview Schedule) by a trained interviewer; and (*e*) continued collection of the psychiatric interview data even if the participant goes off the assigned treatment for any reason, except death or consent withdrawal. Whether the additional information obtained from such a study would justify the time and the expense involved in its collection is a problematic question that is beyond the scope of this article.

References

(1) Cathcart CK, Jones SE, Pumroy CS, Peters GN, Knox SM, Cheek JH. Clinical recognition and management of depression in node negative breast cancer patients treated with tamoxifen. Breast Cancer Res Treat 1993;27:277–81.

(2) Shariff S, Cumming CE, Lees A, Handman M, Cumming DC. Mood disorder in women with early breast cancer taking tamoxifen, an estradiol receptor antagonist. An expected or unexpected effect? Ann N Y Acad Sci 1995;761:365–8.

(3) Anelli TF, Anelli A, Tran KN, Lebwohl DE, Borgen PI. Tamoxifen administration is associated with a high rate of treatment-limiting symptoms in male breast cancer patients. Cancer 1994;74:74–7.

(4) Pluss JL, DiBella NJ. Reversible central nervous system dysfunction due to tamoxifen in a patient with breast cancer. Ann Intern Med 1984;101:652.

(5) Duffy LS, Greenberg DB, Younger J, Ferraro MG. Iatrogenic acute estrogen deficiency and psychiatric syndromes in breast cancer patients. Psychosomatics 1999;40:304–8.

(6) Physician's Desk Reference. 53rd ed. Oradell (NJ): Medical Economics Co.; 1999.

(7) Love RR, Cameron L, Connell BL, Leventhal H. Symptoms associated with tamoxifen treatment in postmenopausal women. Arch Intern Med 1991;151:1842–7.
(8) Fallowfield L, Fleissig A, Edwards R, West A, Powles TJ, Howell A, et al. Tamoxifen for the prevention of breast cancer: psychosocial impact on women participating in two randomized controlled trials. J Clin Oncol 2001;19:1885–92.
(9) Halbreich U. Role of estrogen in postmenopausal depression. Neurology 1997;48(5 Suppl 7):S16–9.
(10) Gregoire AJ, Kumar R, Everitt B, Henderson AF, Studd JW. Transdermal oestrogen for treatment of severe postnatal depression. Lancet 1996;347: 930–3.
(11) Murray D. Oestrogen and postnatal depression. Lancet 1996;347:918–9.
(12) Fisher B, Costantino JP, Wickerham DL, Redmond CK, Kavanah M, Cronin WM, et al. Tamoxifen for the prevention of breast cancer: report from the National Surgical Adjuvant Breast and Bowel Project P-1 Study. J Natl Cancer Inst 1998;90:1371–88.
(13) Fisher B, Costantino J. Highlights of the NSABP Breast Cancer Prevention Trial. Cancer Control 1997;4:78–86.
(14) Ganz PA, Day R, Ware JE Jr, Redmond C, Fisher B. Base-line quality-of-life assessment in the National Surgical Adjuvant Breast and Bowel Project Breast Cancer Prevention Trial. J Natl Cancer Inst 1995;87:1372–82.
(15) Ganz PA, Day R, Costantino JP. Compliance with quality of life data collection in the National Surgical Adjuvant Breast and Bowel Project (NSABP) Breast Cancer Prevention Trial. Stat Med 1998;17:613–22.
(16) Day R, Ganz PA, Costantino JP, Cronin WM, Wickerham DL, Fisher B. Health-related quality of life and tamoxifen in breast cancer prevention: a report from the National Surgical Adjuvant Breast and Bowel Project P-1 Study. J Clin Oncol 1999;17:2659–69.
(17) Radloff LS. The CES-D scale: a self-report depression scale for use in the general population. Applied Psychol Measurement 1977;1:385–401.
(18) Stewart AL, Ware JE. Measuring functioning and well-being: the medical outcomes approach. Durham (NC): Duke University Press; 1992.
(19) Weissman MM, Bruce ML, Leaf PJ, Florio LP, Holzer C. Affective disorders. In: Robins LN, Regier DA, editors. Psychiatric disorders in America: the Epidemiological Catchment Area study. New York (NY): The Free Press; 1991.
(20) Diagnostic and statistical manual of mental disorders: DSM-IV. Washington (DC): American Psychiatric Association; 2000.
(21) Spitzer RL, Endicott J, Robbins E. Research Diagnostic Criteria. New York (NY): Biometrics Research Division, New York State Psychiatric Institute; 1978.
(22) Dohrenwend BP, Shrout PE, Egin G, Mendelsohn FS. Nonspecific psychological distress and other dimensions of psychopathology. Measures for use in the general population. Arch Gen Psychiatry 1980;37:1229–36.
(23) Myers JK, Weissman MM. Use of a self-report symptom scale to detect depression in a community sample. Am J Psychiatry 1980;137:1081–4.
(24) Lin N, Dean A, Ensel WM. Social support, life events, and depression. New York (NY): Academic Press; 1986.
(25) Boyd JH, Weissman MM, Thompson WD, Myers JK. Screening for depression in a community sample. Understanding the discrepancies between depression symptom and diagnostic scales. Arch Gen Psychiatry 1982;39: 1195–200.
(26) Spitzer RL, Endicott J. Schedule for Affective Disorders and Schizophrenia. New York (NY): Biometrics Research Division, New York State Psychiatric Institute; 1978.
(27) Kaelber CT, Moul DE, Farmer ME. Epidemiology of depression. In: Beckham EE, Leber WR, editors. Handbook of depression. 2nd ed. New York (NY): Guilford Press; 1995. p. 3–35.
(28) Keller MB, Lavori PW, Rice J, Coryell W, Hirschfeld RM. The persistent risk of chronicity in recurrent episodes of nonbipolar major depressive disorder: a prospective follow-up. Am J Psychiatry 1986;143:24–8.
(29) Keller MB, Lavori PW, Lewis CE, Klerman GL. Predictors of relapse in major depressive disorder. JAMA 1983;250:3299–304.
(30) Leaf PJ, Livingston MM, Tischler GL, Weissman MM, Holzer CE 3rd, Myers JK. Contact with health professionals for the treatment of psychiatric and emotional problems. Med Care 1985;23:1322–37.
(31) Costantino JP, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J, et al. Validation studies for models projecting the risk of invasive and total breast cancer incidence. J Natl Cancer Inst 1999;91:1541–8.
(32) Fisher B, Costantino J, Redmond C, Poisson R, Bowman D, Conture J, et al. A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor-positive tumors. N Engl J Med 1989;320:479–84.
(33) Fleiss JH. Statistical methods for rates and proportions. 2nd ed. New York (NY): John Wiley & Sons; 1981.
(34) Brown H, Prescott R. Applied mixed models in medicine. New York (NY): John Wiley & Sons; 1999.
(35) Land S, Wieand S, Day R, Ten Have T, Costantino JP, Lang W, et al. Methodological issues in the analysis of quality of life data in clinical trials: illustrations from the National Surgical Adjuvant Breast and Bowel Project (NSABP) Breast Cancer Prevention Trial. In: Mesbah M, Cole B, Lee M, editors. Statistical design, measurement and analysis of health-related quality of life. New York (NY): Kluwer Academic Publishers. In press 2001.

NOTES

Supported by Public Health Service grant NCI-U10CA37377/69974 from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services; by career development award DAMD17–97–1-7058 from the Department of Defense (to R. Day); and in part by an American Cancer Society Clinical Research Professorship (to P. A. Ganz).

We thank Samuel Wieand, Ph.D, Stephanie Land, Ph.D., and Ms. Sheela Goshal of the National Surgical Adjuvant Breast and Bowel Project (NSABP) Biostatistical Center and D. Lawrence Wickerham, M.D., of the NSABP Operations Center for their help in the preparation of this article.

Manuscript received February 20, 2001; revised August 15, 2001; accepted August 31, 2001.

Quality of Life and Tamoxifen in a Breast Cancer Prevention Trial

A Summary of Findings from the NSABP P-1 Study

RICHARD DAY

Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania 15213, USA

ABSTRACT: **This report contains a brief summary of the health-related quality of life findings for 11,064 women taking part in the National Surgical Adjuvant Breast and Bowel Project's P-1 trial. Women taking part in this trial of tamoxifen versus placebo for breast cancer prevention were ≥35 years old and predominantly white, well educated, and middle class, with a strong professional and technical orientation. Key findings included a lack of difference between the tamoxifen and placebo arms with regard to depression, overall physical or mental quality of life, and weight gain. The tamoxifen arm did show consistent increases in vasomotor (hot flashes) and gynecological (vaginal discharge) symptoms, as well as difficulties in certain domains of sexual functioning. It is concluded that an informed discussion with a woman considering tamoxifen therapy should include these points in the risk-benefit discussion.**

KEYWORDS: **quality of life; tamoxifen; breast cancer; prevention**

INTRODUCTION

This is a brief summary of the findings from the health-related quality of life (HRQL) component of the National Surgical Adjuvant Breast and Bowel Project's (NSABP) P-1 trial, a multicenter, double-blinded, placebo-controlled clinical trial designed to evaluate whether 5 years of tamoxifen therapy would reduce the incidence of invasive breast cancer in women at an increased risk for the disease. Detailed descriptions of the rationale, planning, and design of the P-1 study and its HRQL component, as well as specific instruments, are available in separate reports.[1–5]

SUBJECTS AND INSTRUMENTS

This summary focuses on the baseline HRQL examination and the first 36 months of follow-up data on 11,064 women recruited over the first 24 months of the study. The P-1 HRQL questionnaire was composed of the Center for Epidemiological

Address for correspondence: Richard Day, Ph.D., Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, 201 North Craig Street, Suite 350, Pittsburgh, PA 15213. Voice: 412-624-4077; fax: 412-624-9969.
day@nsabp.pitt.edu

Studies–Depression Scale (CES-D), the Medical Outcomes Study (MOS) Short Form (SF-36), the MOS sexual functioning scale, and a symptom checklist (SCL). The questionnaire was to be administered to all participants prior to randomization (baseline), at 3 months, and at each succeeding 6-month examination.

RESULTS

The participants in the P-1 study were predominantly white (96%), well educated (≥65% had some college), married (70%), and professional and technically trained (68.2%) women, who were currently employed (64.9%) and reported a middle to upper-middle class family income (median, $35,000–$49,999).

FIGURE 1 shows the overall proportion and total numbers of women completing the HRQL questionnaire at each examination. It provides a measure of comparative participant adherence with regard to the HRQL questionnaire in the two trial groups. Analysis of sociodemographic and medical variables indicated that participants failing to complete the HRQL questionnaire in each group were similar cohorts of women.

FIGURE 2 shows the proportion of P-1 participants, by group and examination, scoring above the most frequently used clinical cutoff (≥16, i.e., the score above

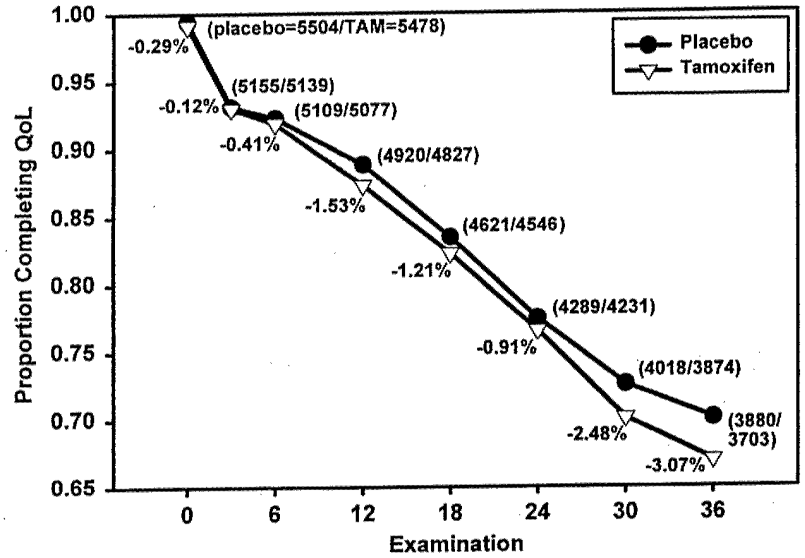


FIGURE 1. Proportion of participants in the tamoxifen (n = 5527) and placebo (n = 5537) groups completing the QoL questionnaire by examination. *Figures in parentheses* are the number of women in the placebo and tamoxifen groups completing the QoL questionnaire. The difference between the tamoxifen and placebo groups is expressed in terms of percent missing QoL data. QoL, quality of life; TAM, tamoxifen.

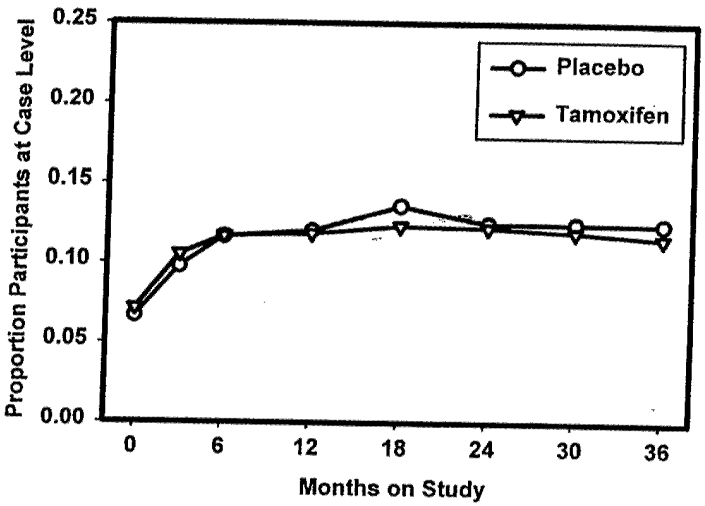


FIGURE 2. Proportion of P-1 participants with CES-D scores at the level of a potential case (≥16) by arm and examination. A CES-D score of ≥16 is the clinical cutoff—that is, it is the score above which depression is considered to be present.

which depression is considered to be present) on the CES-D.[6,7] The youngest age group (35–49 years old) in both trial groups consistently had the highest proportion of members scoring above the clinical cutoff, followed by the 50- to 59-year-old age group. Similar findings with regard to the relationship between the two trial groups emerged from the analysis of the 5-item mental health subscale on the MOS SF-36 (not shown).

The SF-36 results are summarized in FIGURE 3 using the physical and mental component scores (PCS, MCS).[8] Mean PCS declines across the age groups. On follow-up examinations, the tamoxifen group was consistently lower on the PCS only in the 50- to 59-year-old age group (one-sided sign test, $P = 0.065$); however, absolute differences were very small, approximating 1/10th of a standard deviation. No consistent differences emerged on the MCS between the two trial groups.

TABLE 1 provides information on the proportion of women in the tamoxifen and placebo groups reporting symptoms on the SCL at least once during the period that the participants were on treatment—that is, the period excluding baseline, but including the seven follow-up examinations. The five symptoms with the greatest relative difference between the two trial groups are given for each age group and the 10 symptoms with the greatest relative difference are presented for all participants combined.

FIGURE 4 summarizes the information from the five items on the MOS sexual functioning scale. Plate A on FIGURE 4 shows that a greater proportion of participants in the tamoxifen as compared to the placebo group reported being sexually active during the 6 months prior to each follow-up examination. Although apparently consistent, the absolute difference was small (mean = 0.78%) and the findings may have been due

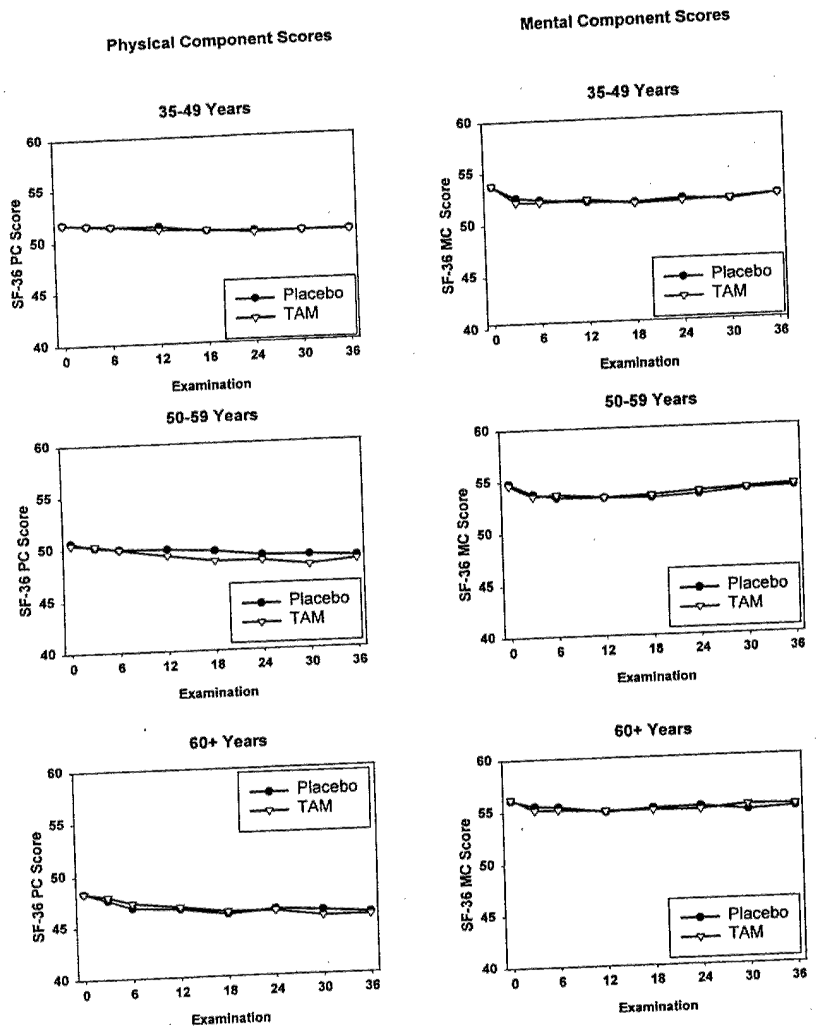


FIGURE 3. Mean scores by age group and examination on SF-36 physical and mental component scores. A higher score represents a better quality of life. TAM, tamoxifen.

TABLE 1. Symptoms reported at least once between 3 and 36 months with the largest relative difference between trial arms

Age group and symptom	Placebo arm proportion (%)	Tamoxifen arm proportion (%)	Relative risk (TAM/placebo)
35–49 Years			
Cold sweats	15.90	22.90	1.44
Vaginal discharge	46.29	62.55	1.35
Pain in intercourse	23.88	31.57	1.32
Night sweats	59.58	74.16	1.24
Hot flashes	65.54	81.28	1.24
50–59 Years			
Cold sweats	16.11	27.00	1.68
Vaginal discharge	32.51	53.47	1.64
Genital itching	36.93	45.24	1.23
Night sweats	62.77	75.88	1.21
Bladder control (laugh)	47.67	56.94	1.19
≥60 Years			
Vaginal bleeding	4.64	10.92	2.35
Vaginal discharge	19.82	45.81	2.31
Genital itching	32.05	40.96	1.28
Hot flashes	51.51	63.59	1.23
Bladder control (laugh)	49.88	56.49	1.13
Overall			
Vaginal discharge	34.13	54.77	1.60
Cold sweats	14.77	21.40	1.45
Genital itching	38.29	47.13	1.23
Night sweats	54.92	66.80	1.22
Hot flashes	65.04	77.66	1.19
Pain in intercourse	24.13	28.19	1.17
Bladder control (laugh)	46.65	52.51	1.13
Bladder control (other)	47.79	52.83	1.11
Weight loss	41.97	44.94	1.07
Vaginal bleeding	21.26	21.96	1.03

to chance. Plates B–E show that a small, but consistently larger percentage of participants in the tamoxifen group reported a definite or serious problem in three of the four specific domains of sexual functioning during the follow-up period.

DISCUSSION

The cohort of women taking part in the P-1 study were not representative of the general population. They were predominantly white, well educated, and middle class, with a strong professional and technical orientation. The initial HRQL findings presented in this report must be assessed within the context of the socioeconomic and cultural characteristics of the P-1 study cohort.

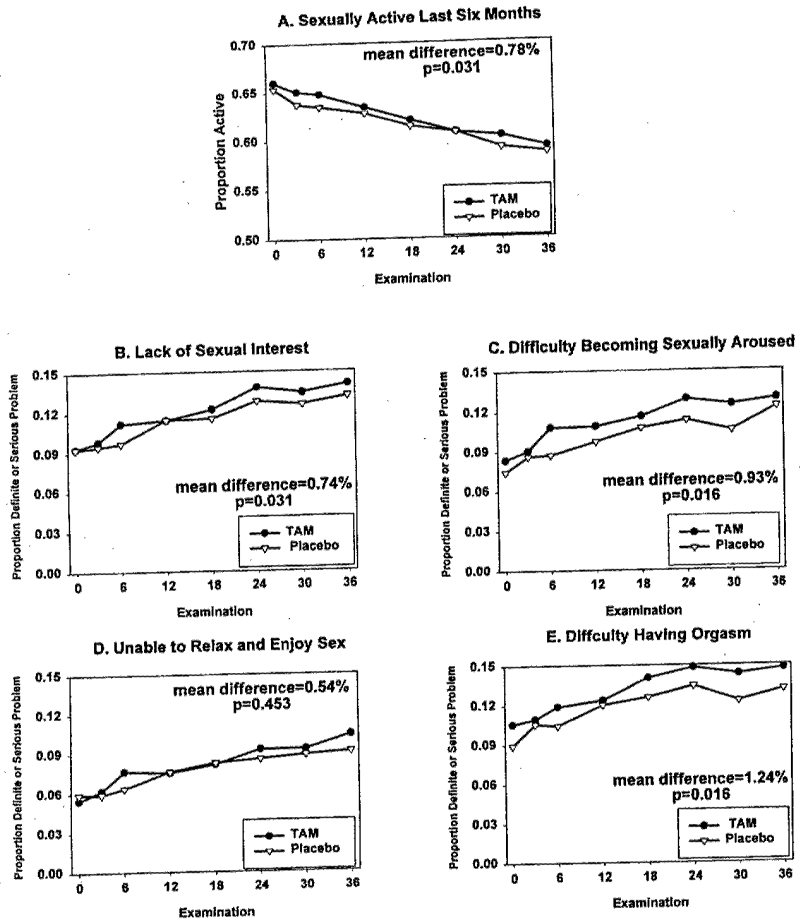


FIGURE 4. Proportion of women in the tamoxifen (TAM) and placebo arms reporting a definite or serious problem in the past 4 weeks on the MOS sexual functioning scale. Plates **B to E** refer only to women who reported being sexually active in the last 6 months.

Although 31.5% of our participants did not complete the 36-month HRQL follow-up examination, we have shown that there is only a small difference in the proportion of nonadherent participants in the tamoxifen and placebo groups and that the nonadherent women in both trial groups are generally similar on key demographic, clinical, and HRQL variables. Given these considerations, it seems unlikely that a maximum difference of 3% in the HRQL follow-up rates between the two groups was sufficient to create a significant bias in our between-group comparisons.

Concern has been expressed regarding the possible relationship between tamoxifen use and the onset of depression.[9–13] Women reporting a history of depressive episodes or of treatment for nervous or mental disorders were not excluded from the trial. If tamoxifen use were associated with the onset of clinically diagnosable depression, we would have expected to see a consistent excess of individuals scoring ≥16 on the CES-D in the tamoxifen group. No such consistent excess was observed.

The MOS SF-36 served in this study as a measure of overall health-related quality of life. We presented data from this instrument in terms of two high-level component scores (PCS and MCS), neither one of which demonstrated any clinically significant differences between the tamoxifen and placebo groups.

The first signs of consistent differences between the tamoxifen and placebo groups were observed in the symptom checklist (SCL). The differences between the trial groups tended to be associated with the types of vasomotor, gynecological, and sexual functioning symptoms previously reported for tamoxifen.[10,14,15]

The data from the MOS sexual functioning scale indicate that relatively small (<4.0%), but consistent differences exist between the two groups with regard to the proportion of women reporting definite or serious problems in at least three specific domains of sexual functioning—sexual interest, arousal, and orgasm. These problems do not appear to be age group–specific. Despite these findings for specific domains of functioning, there is no evidence that these problems result in a reduction in the overall proportion of women in the tamoxifen group who are sexually active.

Based on these data, we would conclude that tamoxifen use is associated with an increase in specific vasomotor, gynecological, and sexual functioning symptoms. At the same time, we did not observe any evidence that overall physical or emotional well-being was significantly affected by these differences in the frequency of symptoms. We also found no evidence on the CES-D or the SF-36 mental health scale for an association in any age group between tamoxifen use and an increase in the proportion of women reporting clinically significant levels of depression.

How should clinicians integrate these research results into decision-making and recommendations to women considering the use of tamoxifen in the setting of prevention? Many symptoms experienced by women who participated in this study are age- and menopause-related and exist independent of the use of tamoxifen. However, several symptoms are substantially more frequent in women using tamoxifen and these include vasomotor symptoms (cold sweats, night sweats, hot flashes), vaginal discharge, and genital itching. Women need to be informed of these possible symptoms. Weight gain and depression, two clinical problems anecdotally associated with tamoxifen treatment in women with breast cancer, were not increased in frequency in this large placebo-controlled trial in healthy women. This is good news that must also be communicated to women.

An informed discussion with a woman considering tamoxifen therapy should include these points in the risk-benefit discussion. Disclosure of likely and unlikely symptoms should prepare a woman for what she might experience and reduce her anxiety or concerns should she embark on preventive therapy. Should a woman experience untoward symptoms after starting tamoxifen treatment, the medication can be discontinued if the symptoms cannot be controlled or her personal assessment of the risks and benefits changes.

REFERENCES

1. FISHER, B., J.P. COSTANTINO, D.L. WICKERHAM *et al.* 1998. Tamoxifen for the prevention of breast cancer: a report from the NSABP P-1 study. J. Natl. Cancer Inst. **90:** 1371–1388.
2. FISHER, B. & J.P. COSTANTINO. 1997. Highlights of the NSABP Breast Cancer Prevention Trial. Cancer Control **4:** 78–86.
3. GANZ, P.A., R. DAY, J.E. WARE *et al.* 1995. Baseline quality-of-life assessment in the National Surgical Adjuvant Breast and Bowel Project Breast Cancer Prevention Trial. J. Natl. Cancer Inst. **87:** 1372–1382.
4. GANZ, P.A., R. DAY & J.P. COSTANTINO. 1998. Compliance with quality of life data collection in the NSABP Breast Cancer Prevention Trial. Stat. Med. **17:** 613–622.
5. DAY, R., P.A. GANZ, J.P. COSTANTINO *et al.* 1999. Health-related quality of life and tamoxifen in breast cancer prevention: a report from the NSABP P-1 study. J. Clin. Oncol. **17:** 2659–2669.
6. RADLOFF, L.S. 1977. The CES-D scale: a self-report depression scale for research in the general population. Appl. Psychol. Meas. **1:** 385–401.
7. ROBERTS, R.E. & S.W. VERNON. 1983. The Center for Epidemiologic Studies Depression Scale: its use in a community sample. Am. J. Psychiatry **140:** 41–46.
8. WARE, J.E., M. KOSINSKI & S.D. KELLER. 1994. SF-36 Physical and Mental Summary Scales: A User's Manual. Third printing revised. The Health Institute, New England Medical Center. Boston.
9. CATHACART, C.K., S.E. JONES, C.S. PUMROY *et al.* 1993. Clinical recognition and management of depression in node negative breast cancer patients treated with tamoxifen. Breast Cancer Res. Treat. **27:** 277–281.
10. LOVE, R.L., L. CAMERON, B.L. CONNELL *et al.* 1991. Symptoms associated with tamoxifen treatment in postmenopausal women. Arch. Intern. Med. **151:** 1842–1847.
11. SHARIFF, S., C.E. CUMMING, A. LEES *et al.* 1995. Mood disorder in women with early breast cancer taking tamoxifen, an estradiol receptor antagonist: an unexpected effect? Ann. N.Y. Acad. Sci. **761:** 365–368.
12. MOREDO ANELLI, T., A. ANELLI, K.N. TRAN *et al.* 1994. Tamoxifen administration is associated with a high rate of treatment-limiting symptoms in male breast cancer patients. Cancer **74:** 74–77.
13. PLUSS, J.L. & N.J. DIBELLA. 1984. Reversible central nervous system dysfunction due to tamoxifen in a patient with breast cancer. Ann. Intern. Med. **101:** 652.
14. FISHER, B., J. DIGNAM, J. BRYANT *et al.* 1996. Five versus more than five years of tamoxifen therapy for breast cancer patients with negative lymph nodes and estrogen receptor–positive tumors. J. Natl. Cancer Inst. **88:** 1529–1542.
15. FISHER, B., J.P. COSTANTINO, C. REDMOND *et al.* 1989. A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen receptor–positive tumors. N. Engl. J. Med. **320:** 479–484.

Selective Estrogen Receptor Modulators and Cardiovascular Disease

Introduction

DAVID J. GORDON

Division of Heart and Vascular Diseases, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

The five papers presented in this section review the evidence for favorable cardiovascular effects of estrogen and compare these effects with those of the new selective estrogen receptor modulators (SERMs).[1–5]

Dr. Herrington has reviewed the many lines of evidence suggesting that oral administration of estrogen may be beneficial in the prevention and treatment of coronary heart disease (CHD) in postmenopausal women.[1] Estrogen has favorable effects on cardiovascular risk factors, including high-density (HDL) and low-density (LDL) lipoprotein cholesterol and fibrinogen, and on measures of vascular endothelial function. Coadministration of medroxyprogesterone acetate (MPA) to protect against endometrial proliferation and neoplasia only partially offsets these putative benefits. CHD incidence and mortality in women generally lag 15 years behind men and are relatively uncommon before menopause. Prospective observational epidemiologic studies in postmenopausal women have shown an association of hormone replacement therapy (HRT), usually combined equine estrogens (CEE) with or without MPA, with marked reduction in risk of CHD and in mortality. However, randomized clinical trials of HRT completed to date have failed to show significant benefit of HRT on clinical CHD events or on the progression of the underlying coronary atherosclerotic lesions and suggest adverse effects on the early incidence of myocardial infarction (MI) and on the incidence of venous thromboembolism (VTE).[1]

The recent advent of the SERMs, which mimic some of the effects of estrogen while antagonizing others, offers the possibility to attain the potential cardiovascular benefits of estrogen therapy without its potential adverse cardiovascular effects or its attendant increased risk of carcinoma of the uterus and breast. In fact, raloxifene, one of the most widely used SERMs, acts as an antiestrogen in the uterus and breast, and does not require coadministration of a progestin in gynecologically intact women.

The articles by Dr. Walsh,[2] Drs. Blum and Cannon,[3] and Dr. Cushman[4] compare the effects of raloxifene and other SERMs with conventional HRT on lipoproteins and on measures of vascular endothelial function, inflammation, and hemostasis.

Address for correspondence: David J. Gordon, M.D., Ph.D., M.P.H., National Heart, Lung, and Blood Institute, Division of Heart and Vascular Diseases, 2 Rockledge Center, Suite 9044, 6701 Rockledge Drive, Bethesda, MD 20892-7940. Voice: 301-435-0564; fax: 301-480-1335. gordond@nhlbi.nih.gov

METHODOLOGICAL ISSUES IN THE ANALYSIS OF QUALITY OF LIFE DATA IN CLINICAL TRIALS: ILLUSTRATIONS FROM THE NATIONAL SURGICAL ADJUVANT BREAST AND BOWEL PROJECT (NSABP) BREAST CANCER PREVENTION TRIAL

STEPHANIE LAND[1], SAMUEL WIEAND[1], RICHARD DAY[1], TOM TEN HAVE[2], JOSEPH P. COSTANTINO[1], WEI LANG[3], AND PATRICIA A. GANZ[4]

[1] University of Pittsburgh and the National Surgical Breast and Bowel Project, [2] University of Pennsylvania School of Medicine, [3] Wake Forest University School of Medicine, [4] UCLA Schools of Medicine and Public Health and the Jonsson Comprehensive Cancer Center

We present two Quality of Life (QOL) endpoints collected in conjunction with the recently completed Breast Cancer Prevention Trial (BCPT) performed by the National Surgical Adjuvant Breast and Bowel Project. The analyses of these endpoints (depression and hot flashes) indicate the importance of randomization and give some insight about the impact of missing data in a large randomized trial.

1. Introduction

Quality of life (QOL) assessments have been increasingly included as secondary or primary endpoints in clinical trials (Tannock *et al.*, 1996; Moinpour *et al.*, 1998). The impetus for doing so comes from a desire to obtain patient-rated evaluations of treatments, especially in circumstances in which treatments have substantially differing toxicities or in which survival outcomes are not expected to be different (Ganz, 1994 a and b). Under such circumstances, an evaluation of the morbidity of treatment from the patient's or participant's perspective may in fact be the most important endpoint. Although there is now a wide range of psychometrically validated scales for the measurement of QOL in clinical trials (Cella and Bonomi, 1995), there are considerable challenges to the implementation and collection of QOL data in these studies (Bernhard *et al.*, 1998a), as well as equally formidable statistical and analytical concerns (Bernhard and Gelber, 1998b). In this paper, we provide examples from the recently completed NSABP Breast Cancer Prevention Trial (BCPT) to highlight challenges that can arise in the analyses of such data, specifically focusing on the importance of randomization and the issue of missing data and its potential to affect the interpretation of QOL outcomes.

2. Background

The BCPT was a double-blinded, placebo-controlled clinical trial that was open for accrual from June 1, 1992, through September 30, 1997. During this interval 13,338 women at high risk for breast cancer were randomly assigned to receive either 20 mg/day of tamoxifen or placebo for a duration of five years. The primary objective of the trial was to determine if tamoxifen therapy would reduce the risk of breast cancer among women. Secondary objectives related to the full benefit/risk profile of tamoxifen use in healthy women. Participants in the trial were screened for breast cancer at six-month intervals by clinical breast examination and at yearly intervals by bilateral mammography. At each screening, visit participants were also evaluated for several other endpoints including heart disease, fractures, thromboembolic disease, and endometrial cancer. Heart disease and fractures were included because it was theorized that tamoxifen might also reduce the risk of these problems. Thromboembolic disease and endometrial cancer were included

because these were known side effects associated with tamoxifen therapy. As an additional means to monitor the safety of treatment in the trial, the incidence of all invasive cancers and the occurrence of all deaths were also included as endpoints.

The results of the BCPT have been reported (Fisher *et al.*, 1998), as has a study of the risk-benefit ratio for tamoxifen (Gail *et al.*, 1999). During follow-up, 175 participants randomized to receive placebo developed invasive breast cancer compared to only 89 randomized to receive tamoxifen, indicating an estimated 50% reduction in the risk of breast cancer from the use of tamoxifen. Other major findings include the detection of a preventive effect on osteoporotic fractures, no effect on heart disease, and a confirmation of the known side effects of endometrial cancer and thromboembolic disease. These overall findings will not be discussed in this manuscript, as we wish to focus on issues that relate to QOL studies.

Because the participants in this trial were healthy women, the monitoring of their QOL during the intervention was of particular importance. Thus, the NSABP included a concurrent QOL study designed to describe side effects of tamoxifen, to examine the relationship between side effects and QOL, to compare the side effects and QOL in placebo and treated subjects, and to examine the effects of symptoms on compliance with study medication. The BCPT QOL questionnaire was a 104-item battery that included four instruments: the Center for Epidemiologic Studies Depression Scale (CES-D) (Radloff, 1997); the Medical Outcomes Study 36-Item Short Form (SF-36) (Ware *et al.*, 1994); a symptom checklist based on the Postmenopausal Estrogen Progesterone Intervention (Shumaker S., personal communication) specifically adapted for the BCPT trial; and the Medical Outcomes Study Sexual Problems Scale (Sherbourne, 1992). These instruments were selected because of their psychometric characteristics and validity, the availability of normative data in healthy women, and ease of self-administration. The latter was particularly important because the trial was conducted at several hundred clinical centers throughout North America and the battery of questions we asked was completed on multiple occasions in conjunction with study visits. The QOL assessment was scheduled to occur at baseline before administration of the study medication and at every clinical visit during the five years after randomization (at three months, at six months, and every six months thereafter). However, the trial was unblinded on March 31, 1998, following an interim analysis that showed a dramatic reduction in the incidence of breast cancer among the participants who received tamoxifen. The QOL follow-up was terminated at that time due to the potential loss of the control arm. In this manuscript, as in our prior analyses of QOL data from the trial (Day *et al.*, 1999, Ganz *et al.*, 1998), we use QOL data available on participants who were recruited to the trial during the first two years of the study (June 1, 1992 to May 31, 1994) as all of these women would have been expected to have 36 months of completed follow-up data at the time the study was terminated. The sample includes 11,064 women who represent 82.6% of the total accrual to the BCPT. We use only their first three years of follow-up.

3. The Effect of Tamoxifen on Depression

When the BCPT began, there was considerable concern that tamoxifen therapy might be associated with the development of depressed mood in women with breast cancer. Although Love et al. (1991) did not find such an effect when reporting symptoms associated with tamoxifen treatment in a randomized trial in postmenopausal women with breast cancer, several researchers subsequently reported results suggesting that administration of tamoxifen might lead to depression in some breast cancer patients (Cathacart *et al.*, 1993; Shariff *et al.*, 1995; Moredo *et al.*, 1994). The latter studies were relatively small (fewer than 400 patients) and none had a placebo comparison group. However, there was a potential scientific rationale for tamoxifen's association with depression. Estrogen had been shown to have a beneficial effect on mood in postmenopausal

women (Halbreich, 1997; Gregoine, *et al.*, 1996), and it was considered plausible that tamoxifen might negate these positive effects of estrogen. Thus, careful measurement of depression, including a screening instrument to identify potential cases of depression, was important in the design of the BCPT QOL study.

The primary instrument used in the BCPT to study the change in depression level over time was the CES-D, a self-administered questionnaire (20 questions) that screens for depressive symptoms over the seven days prior to administration (Radloff, 1977). A participant's score is the sum of the responses for the 20 questions and can range from 0 (no depressive symptoms) to 60 (maximum depressive symptoms). The instrument is widely used because it is easy to administer and has excellent population-based normative data (Myers and Weissman, 1980; Roberts and Vernon, 1983; Boyd *et al.*, 1982). To assess the validity of the CES-D in the BCPT sample, we compared the baseline CES-D scores of BCPT participants with ten medical history items related to mental health that had been obtained at entry to the trial (Table 1). The first three items were obtained in the context of questions about diagnosed medical problems, although we did not verify that there had been a recorded diagnosis. There is nearly a linear relationship between the number of positives from the participant's mental health history and the CES-D score ($p<0.0001$), providing considerable reassurance that the CES-D score from this study sample was highly associated with a clinical mental health history. Similarly, the association between the mean CES-D score and the three depression-related items "ever had depression," "ever took antidepressants" (either item 4 or 6 positive), or "any two years depressed or sad", showed an increasing relationship between the CES-D score and the number of positives ($p<.001$). In addition, the baseline CES-D scores were well balanced across placebo and tamoxifen treatment assignment (Table 2). Cut-off points used in the table are somewhat arbitrary, although a cut-off of 16 is commonly used as the minimum for classifying a person as depressed (Myers and Weissman, 1980; Roberts and Vernon, 1983; Boyd *et al.*, 1982) and Lyness *et al.* (1997) used the cutoff of 22 when screening for major depression.

Table 1
BCPT Participant History Mental Health Items Obtained at Entry to the BCPT

Item	% Yes
Ever had depression	15
Ever had nervous or emotional disorder	3
Ever had psychiatric problems	1
Current antidepressants	6
Current tranquilizers	16
Previous antidepressants	4
Previous tranquilizers	15
Two weeks sad, blue, depressed, disinterested	17
Any two years depressed or sad	9
Depressed or sad most of past year	5

Table 2. Baseline CES-D Scores

Score	Placebo (%)	Tamoxifen (%)

0-10	85.5	83.9
11-15	7.8	9.0
16-21	3.8	4.1
22-60	2.8	3.0

In Figure 1, we present the mean CES-D scores by visit and treatment arm during the BCPT. The observed increase of depression among participants receiving tamoxifen is slightly less than the observed increase among participants receiving placebo, although the difference is not significant (p=0.24). Thus, the increase in the depression score during the first six months of the trial does not appear to be related to the administration of tamoxifen. It is noteworthy that the dramatic increase in scores at months 3 and 6 would almost certainly have been attributed to tamoxifen had there not been a placebo arm. This illustrates the danger of trying to establish a cause-and-effect relationship in a non-randomized setting.

FIGURE 1 ABOUT HERE.

We do not know why the CES-D depression scores increased for participants on both arms of the study (placebo and tamoxifen). It is possible that symptoms of worry and depression increased due to the controversy surrounding this trial, the fear and uncertainty of taking either placebo or active agent, an increased awareness of breast cancer risk, or a concern over potential therapy side effects. Alternatively, the raised scores might be partially attributed to "nocebo effect" (Hahn, 1997): if an individual fears or believes that a side effect may occur from a medication, he or she will report it. (As will be shown later, participants receiving placebo also reported an increase in hot flashes, but not at the same significant rate as the participants on tamoxifen.) Since neither group of women knew which pill they were taking, they may have reported increased symptoms because they feared the potential medication side effects described to them as part of the consent process. A third possibility is that the baseline scores were artificially low and the subsequent increase reflected a regression to the mean. We do not believe the baseline scores are much lower than would be expected for the educated, socioeconomically advantaged population in the trial. However, to the extent that the scores were artificially low, it could be either that women were less likely to enter the trial when they were experiencing depressive symptoms, or that they would under-report for fear of jeopardizing their inclusion in the trial. In any case, the phenomenon of an early increase in depressive symptoms appears to be independent of tamoxifen use.

However, we were concerned that there might be a treatment effect in the subset of subjects at higher risk of depression. Because 93% of the participants had baseline CES-D scores <16, and 85% had scores <11, such an effect might not be apparent in an analysis based on the entire population. To explore this possibility we divided the women into four groups of risk: zero, one to two, three to five, and six to ten "yes" responses to the mental health items listed in Table 1. There was no difference observed between tamoxifen and placebo participants in any of the four groups. Results were similar when the baseline CES-D score was used to create risk groups (CES-D scores from 0 to 11; 12 to 15; 16 to 21; or 22 or more). There was a suggestion that tamoxifen is beneficial in the high-risk group (p=0.04), although this is likely to be a statistical artifact.

The problem of missing data is common in clinical trials that assess QOL (Bernhard *et al.*, 1998a). In the BCPT, this was exacerbated by the fact that the clinical centers were not required to collect QOL data when a participant went off the study medication. As will be seen, this led to a substantial problem of non-random missing data. Only 82 participants did not fill out the CES-D form at entry (an extremely low rate of missing baseline data), and these participants were excluded from subsequent analyses. (Questionnaires that were partially completed are

considered missing in this report.) However, of the possible 76,874 post-entry forms that 10,982 remaining participants were expected to submit during the three-year period, 13,752 (18%) were missing. At the end of the third year, slightly more than 30% were missing and participants who received tamoxifen were more likely to have missing data (33% versus 30% missing, $p<0.001$). The first three rows in Table 3 present the number and percent of missing forms preceded by a protocol-specified event (such as second primary cancer, deep-vein thrombosis, ischemic heart disease, or death); missing forms preceded by early termination of therapy for a reason not specified by the protocol; and missing forms preceded by consent withdrawal by the participant; the fourth row of this table shows the number of forms that were missing when the participant was still receiving therapy. Figure 2 displays the percent of missing forms in four groups based on baseline CES-D scores. Participants who began with an elevated CES-D score were more likely to have missing data ($p<0.001$ at three years) .

INSERT FIGURE 2 AND TABLE 3 ABOUT HERE.

The average of the CES-D scores immediately preceding a missing score was higher than the average of the CES-D scores immediately preceding an observed score (Table 4), which raised the possibility that missing scores would have been higher than concurrently observed scores. The differences were almost identical in the tamoxifen and placebo arms, indicating that while the missing data might result in an underestimate of depression, the bias would be the same in both arms. When we considered other functions of preceding scores, we found that none had a stronger association with missing scores than did the immediately preceding score. In particular, the slopes between two scores preceding a missing score were not significantly different from slopes between two scores preceding an observed score. Therefore we considered some simple imputations based on the scores immediately preceding the missing scores.

INSERT TABLE 4 ABOUT HERE.

In discussing the imputation methods, we will use the following notation. The baseline and seven post-entry CES-D scores for the *j*th individual participant will be represented by the vector $\underset{\sim}{x}_j = (x_{0j}, x_{1j}, x_{2j}, x_{3j}, x_{4j}, x_{5j}, x_{6j}, x_{7j})$, where "missing" is a possible value for the CES-D score. Let $\overline{x}_{i.}^{T}$ ($\overline{x}_{i.}^{P}$) be the average CES-D among tamoxifen (placebo) participants with an observed CES-D score at the *i*th visit. We define a new set of vectors by $\underset{\sim}{x}_i^I = (x_{0i}, x_{1i}^I, x_{2i}^I, x_{3i}^I, x_{4i}^I, x_{5i}^I, x_{6i}^I, x_{7i}^I)$, where $x_{ij}^I = x_{ij}$ if x_{ij} is observed. If x_{ij} is missing, $x_{ij}^I = x_{(i-1)j}^I + \overline{x}_{i.}^{T} - \overline{x}_{(i-1).}^{T}$ for a tamoxifen participant and $x_{ij}^I = x_{(i-1)j}^I + \overline{x}_{i.}^{P} - \overline{x}_{(i-1).}^{P}$ for a placebo participant, where the imputation begins with x_{1i}^I then x_{2i}^I and so forth. The mean CES-D curves are slightly higher than in Figure 1 (where no imputation is involved), but the differences between the two curves remain nearly identical to the differences seen in Figure 1.

Although Table 4 suggests that the imputed values defined above would be appropriate for replacing missing values, we cannot rule out the possibility that the missing values mask a greater increase in depression for tamoxifen participants than for placebo participants. For example, there might have been a subset of tamoxifen participants who became depressed as a result of the treatment and dropped out before this effect could be observed. We do not have data available to verify that this is not the case. In order to see just how great a differential (by treatment) would

have been required to change the interpretation of the data, we performed three sensitivity analyses.

For the first sensitivity analysis, we imputed missing values as defined above, but for every missing value of a tamoxifen participant we added 0.5 units to the imputed value. The resultant mean values of CES-D at each assessment are almost the same between treatment arms. This is somewhat reassuring, since adding .5 units to each missing CES-D score for tamoxifen participants and none for placebo participants is extreme. As Table 3 indicated, the status of the participants with missing forms was similar on both arms. In instances in which institutions reported the reason participants went off study, only 3% reported depression as the reason for doing so.

The second sensitivity analysis was based on a partitioning of missing questionnaires into those that were missing for a variety of non-treatment-related reasons and those that were missing for treatment-related reasons. Specifically, we assumed that if m questionnaires were missing (at a particular assessment time) in the placebo arm, and $m+x$ questionnaires were missing in the tamoxifen arm, then some fraction of the x questionnaires might be attributable to excess depression caused by tamoxifen. We calculated treatment group means (at each assessment time) as if some fraction r (for various candidate values of r) of the missing tamoxifen scores were replaced with the mean of all observations at that assessment that were at least 16, since these represent severe depressive symptoms. The remaining missing values in both arms were replaced with the mean of all observations at that assessment. At $r=1/2$, the curves of imputed CES-D for the two treatment groups overlapped [not shown]. That is, there did not appear to be a tamoxifen-related increase in CES-D unless greater than half of the excess missing questionnaires were assumed to coincide with severe depressive symptoms.

All of the analyses shown above were also carried out for a binary outcome of severe depressive symptoms, defined as any CES-D score $\geq$16. In Figure 3.A, we plot the proportion of values classified as a "yes" as a function of time and again find no tamoxifen effect. Imputation of the missing values using preceding scores had minimal impact on our findings. For a sensitivity analysis, we performed the imputation with the additional assumption that 3.2% of the missing tamoxifen CES-D forms had a score $\geq$ 16, even though the prior score was <16. This would be roughly equivalent to assuming that all the tamoxifen participants who reported depression before dropping out of the study subsequently had a score $\geq$ 16, while none of the placebo participants reporting depression before dropping out had a score exceeding 15. The sensitivity analysis, presented in Figure 3.B, indicates that under this fairly extreme assumption about the drop-outs, the two curves would essentially overlap.

INSERT FIGURES 3A AND 3B HERE.

As a final step in the sensitivity analysis, we considered a model-based method that adjusts for drop-out related to observed and unobserved CES-D outcomes through subject-level random effects. This approach, which may be used to adjust for other covariates, has been presented previously in other randomized trial contexts for continuous data (Schlucter, 1992; DeGruttola and Tu, 1994) and for binary data (Ten Have *et al.*, 1998), and in a cohort study context for ordinal data (Ten Have *et al.*, 2000). More specifically, we fitted an ordinal logistic model with random effects to the CES-D outcome data. The CES-D score was categorized as in Table 2. The models make the proportional odds assumption, that is, the odds ratio specified for a given cut-point of the ordinal CES-D scale is the same as the odds ratio specified for every other cut-point. This approach is not designed for intermittent missingness. Therefore, any participant's data subsequent to a missing form was deleted for the purpose of this analysis. The model comprised three components consisting of different covariate effects but sharing the same subject-level random effect structure. The first was an ordinal CES-D outcome component with treatment arm and time (7 degrees of freedom) as main effects, and their interaction (7 degrees of freedom). The

second and third model components corresponded to separate discrete survival time logistic specifications for non-protocol and protocol specified drop-out. Each of these drop-out components included main effect covariates corresponding to treatment arm and time.

We present results based on two versions of each of these drop-out components. The first version includes as covariates the CES-D outcome before drop-out and its interaction with treatment arm and type of dropout (protocol vs. non-protocol). In the second version, each of these drop-out components excludes the CES-D outcome and its interactions. The ensuing results are based on these model specifications without baseline covariates. Including baseline age did not alter the results. The subject-level random effect structure shared by the CES-D and drop-out components induces a relationship between the CES-D observed and unobserved outcomes and the risk of drop-out. The magnitude of this relationship is characterized by the specification of separate variance components of the random effect for each of the three components in the model. Separate large variance components for the outcome component and for a drop-out component indicate a strong relationship between outcome and the respective form of drop-out. For comparison, we also present results based on the assumption that drop-out is *missing at random (MAR)*. That is, drop-out is conditionally independent of the unobserved CES-D outcomes, conditioned on all observed data (Little, 1995). In summary, we have used these three models: 1) the random effects logistic model without a drop-out component, under the assumption that drop-out is missing at random (naïve model); 2) the random effects logistic model augmented with a discrete time survival logistic model for drop-out, which shares a random effect with the ordinal CES-D outcome (Joint 1 model); and 3) model #2 with the last observed CES-D outcome added as a covariate (Joint 2 model).

The likelihood ratio test of treatment arm differences in change across time (7 degrees of freedom) was not significant (p=.14). As Table 5 suggests, this result was robust with respect to the drop-out assumptions (e.g., MAR). More specifically, the estimates of the log treatment odds ratio at baseline and corresponding treatment-time interaction terms at each follow-up time differ very little across the three models. To evaluate the strength of the relationship between outcome and drop-out, we present the variance components of the random effect shared by the three components (outcome, two drop-out types: non-protocol- and protocol-defined) two of the models, Joint 1 and Joint, 2 in Table 6. Note that the naive model only has the outcome component and therefore only one variance component. Table 6 shows that neither of the drop-out components in Joint 1 and Joint 2 models is related to the outcome through a random effect. This lack of relationship between outcome and drop-out is consistent with the fact that the log odds ratio estimates in Table 5 are very stable between the naïve and joint models. This suggests that the naïve random effects model accommodates the relationship between outcome and protocol-defined drop-out. That is, the MAR relationship under the naïve model characterizes the type of relationship between drop-out and outcome represented by the joint models. Of course, it may be that a different relationship exists that is not characterized by either the joint or naïve models.

INSERT TABLES 5 AND 6 ABOUT HERE.

In summary, our study data indicate that tamoxifen does not influence depressive symptoms among women who are at high risk for breast cancer, and there is no indication that missing data masked an effect. It appeared that the missing data did result in slight underestimates of the CES-D scores, which were increased following imputation.

4. Strategies for the Evaluation of Missing Data: Hot Flashes

Although tamoxifen did not appear to influence the CES-D score in this study, it clearly was associated with other symptoms. Numerous studies have shown that tamoxifen increased the number and severity of hot flashes in women being treated for cancer, and this effect was also seen in the high-risk women participating in the BCPT (Day *et al.*, 1999). Hot flash was the most commonly reported symptom on either arm of the BCPT.

In Figure 4 (solid lines), we present the score reported by these women for hot flashes at each cycle by treatment (possible values ranged from 0=none to 4=extreme). There is a clear increase in this symptom associated with tamoxifen throughout the study. (Note that participants taking placebo also report an increase in mean hot flash score, although this increase is not as great as for those taking tamoxifen. This may be another example of the nocebo effect.) Differences in hot flashes due to treatment are highly significant ($p<.001$) at every visit. However, when hot flash scores immediately preceding a missing value were compared to the scores immediately preceding an observed value (Table 7), there was a differential effect according to treatment. We again did an imputation in which missing values were replaced by the prior score adjusted for the mean for the visit (as described previously for the CES-D analyses). There is still clear evidence of a tamoxifen effect (dashed lines in Figure 4), but the values for the tamoxifen curve are slightly lower than when the missing values are omitted, while the values for the placebo curve remain nearly unchanged, indicating that we might be slightly overestimating the treatment effect if we ignore missing values. For example, the difference in average scores is .30 at three years when missing data are ignored versus .26 following the imputation.

INSERT FIGURE 4

Table 7. Average hot flash score prior to missing versus observed scores

	Missing subsequent questionnaire	Observed subsequent questionnaire
Placebo	0.87	0.77
Tamoxifen	1.12	1.16

An alternative analysis of these data based on the informative drop-out model used for the CES-D revealed a significant difference between the treatment arms with respect to change at each follow-up time ($p<.001$). As with the CES-D non-significant treatment difference, this significant result was robust with respect to drop-out assumptions under the random effects ordinal logistic model. The logistic model requires the assumption that the relationship between symptoms and drop-out risk is in the same direction in both the placebo and tamoxifen groups and over time. As Table 8 indicates, this assumption did not hold for the hot flash data. Hence, we were unable to adjust for the observed drop-out pattern to obtain valid estimates of the treatment effect.

INSERT TABLE 8 ABOUT HERE

5. Conclusions

Several points became clear in the analysis of the CES-D data. Perhaps the most important is that one would be likely to conclude that tamoxifen increased depressive symptoms if all the participants had received tamoxifen, as this would appear to be the most likely cause of the immediate increase in depressive symptoms. However, the randomization allowed us to see that the effect increase was comparable when participants received placebo, ruling out tamoxifen as the cause. The fact that the prior scores associated with missing values were elevated in both arms

indicated that the degree of depressive symptoms might have been underestimated slightly on both arms. However the elevation was the same in both arms, which made it unlikely that there was a differential drop-out effect by treatment. This partially explains why imputation analyses still led to the conclusion that tamoxifen did not result in increased depressive symptoms. Sensitivity analyses indicated that even if there were a fairly substantial treatment related difference in the depressive symptoms among the drop-outs, accounting for this differential effect would not change the conclusion that the depressive symptoms were not treatment related.

The situation was slightly different for the hot flash outcome. There was a clear substantial effect of tamoxifen on the incidence and severity of hot flashes. Furthermore, there was evidence of a differential drop-out effect by treatment. Imputation indicated that this resulted in a small overestimate of treatment effect. The rather unusual relationship between drop-outs and treatment presented in Table 8 would require fairly flexible models if one were to estimate and make inference regarding the effect. In future methodology studies, we will address ways to handle this drop-out pattern.

ACKNOWLEGEMENTS

We would like to give special thanks to the participants in the BCPT and the data coordinators at the participating sites, without whom this QOL study would not have been possible. We also wish to note that this work was supported in part by NIH/NCI Grant U10 CA69974 and DAMD grant 17-97-1-7058. We thank Barbara C. Good, Ph.D., for excellent editorial comments. In addition, we thank Maria Harper, Ph.D., and Ginny Mehalik, M.A., for editorial assistance with the manuscript.

REFERENCES

Bernhard, J., Cella, D.F., Coates, A.S., et al. (1998a), "Missing Quality of Life Data in Cancer Clinical Trials: Serious Problems and Challenges," *Statistics in Medicine*, 17, 517-532.

Bernhard, J. and Gelber, R.D., editors (1998b), "Workshop on Missing Data in Quality of Life Research in Clinical Trials: Practical and Methodological Issues," *Statistics in Medicine*, 17, 511-796.

Boyd, J.H., Weissman, M.M., Thompson, W.D., and Myers, J.K. (1982), "Screening for Depression in a Community Sample," *Archives of General Psychiatry*, 39, 1195-1200.

Cathacart, C.K., Jones, S.E., Pumroy, C.S., et al. (1993), "Clinical Recognition and Management of Depression in Node Negative Breast Cancer Patients Treated with Tamoxifen," *Breast Cancer Research and Treatment*, 27, 277-281.

Cella, D.F., and Bonomi, A.E. (1995), "Measuring Quality of Life: 1995 Update," *Oncology*, 9 (11-suppl), 47-60.

Day, R., Ganz, P.A., Costantino, J.P., et al. (1999), "Health-Related Quality of Life and Tamoxifen in Breast Cancer Prevention: A Report from the National Surgical Adjuvant Breast and Bowel Project P-1 Study," *Journal of Clinical Oncology*, 17, 2659-2669.

DeGruttola, V., and Tu, X.M. (1994), "Modelling Progression of CD4-Lymphocyte Count and its Relationship to Survival Time," *Biometrics*, 50, 1003-1014.

Fisher, B., Costantino, J.P., Wickerham, D.L., et al. (1998), "Tamoxifen for Prevention of Breast Cancer: Report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study," *Journal of the National Cancer Institute*, 90, 1371-1388.

Gail, M.H., Costantino, J.P., Bryant, J., et al. (1999), "Weighing the Risks and Benefits of Tamoxifen Treatment for Preventing Breast Cancer," *Journal of the National Cancer Institute*, 91, 1829-1846.

Ganz, P.A. (1994a), "Long Range Impact of Clinical Trial Interventions on Quality of Life," *Cancer*, 74 (9-suppl), 2620-2624.

Ganz, P.A. (1994b), "Quality of Life Measures in Cancer Chemotherapy: Methodology and Implications," *PharmacoEconomics*, 5, 376-388.

Ganz, P.A., Day, R., Ware, J.E., et al. (1995). "Base-Line Quality-of-Life Assessment in the National Surgical Adjuvant Breast and Bowel Project Breast Cancer Prevention Trial," *Journal of the National Cancer Institute*, 87, 1372-1382.

Ganz, P.A., Day, R., Costantino, J. (1998), "Compliance with Quality-of-Life Data Collection in the National Surgical Adjuvant Breast and Bowel Project (NSABP) Breast Cancer Prevention Trial," *Statistics in Medicine*, 17, 613-622.

Gregoine A.J., Kumar, R., Everitt, B., et al. (1996), "Transdermal Oestrogen for Treatment of Severe Postnatal Depression," *Lancet*, 347, 930-933.

Hahn, R. A. (1997). "The Nocebo Phenomenon: Concept, Evidence, and Implications for Public Health"

Halbreich, U. (1997). "Role of Estrogen in Postmenopausal Depression," *Neurology*, 48, 1718-1729.

Little, R.J.A. (1995). "Modeling the Drop-Out Mechanism in Repeated Measures Studies," *Journal of the American Statistical Association*, 90, 1112-1121.

Love, R.R.L., Cameron, L., Connell, B.L., et al. (1991), "Symptoms Associated with Tamoxifen Treatment in Postmenopausal Women," *Archives of Internal Medicine*, 151, 1842-1847.

Lyness, J.M., Noel, T.M., Cox, C., et al. (1997), "Screening for Depression in Elderly Primary Care Patients: A Comparison of the Center for Epidemiologic Studies-Depression Scale and the Geriatric Depression Scale," *Archives of Internal Medicine*, 157, 449-454.

Moinpour, C.M., Savage, M.J., Troxel, A., et al. (1998), "Quality of Life in Advanced Prostate Cancer: Results of a Randomized Therapeutic Trial," *Journal of the National Cancer Institute*, 90, 1537-1544.

Moredo, A.T., Anelli, A., Tran, K.N., et al. (1994), "Tamoxifen Administration is Associated with a High Rate of Treatment Limiting Symptoms in Male Breast Cancer Patients," *Cancer*, 74, 74-77.

Myers, J.K., and Weissman, M.M. (1980). "Use of a Self-Report Symptom Scale to Detect Depression in a Community Sample," *American Journal of Psychiatry*, 137, 1081-1084.

Radloff, L.S. (1977), "The CES-D Scale: A Self-Report Depression Scale for Research in the General Population," *Applied Psychological Measurement*, 3, 385-401.

Roberts, R.E., and Vernon, S.W. (1983), "The Center for Epidemiologic Studies Depression Scale: Its Use in a Community Sample," *American Journal of Psychiatry*, 140, 41-46.

Schluchter, M. (1992), "Methods for the Analysis of Informatively Censored Longitudinal Data," *Statistics in Medicine*, 11, 1861-1870.

Shariff, S., Cumming, C.E., Lees, A., et al. (1995), "Mood Disorder in Women with Early Breast Cancer Taking Tamoxifen, an Estradiol Receptor Antagonist: An Unexpected Effect?" *Annals of the New York Academy of Sciences*, 761, 365-368.

Sherbourne, C.D. (1992), In: Stewart A.L., Ware, J.E. Editors. *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach,* Durham (NC): Duke University Press, 194-204.

Tannock, I.F., Osoba, D., Stockler, M.R., et al. (1996), "Chemotherapy with Mitoxantrone plus Prednisone or Prednisone Alone for Symptomatic Hormone-Resistant Prostate Cancer: A Canadian Randomized Trial with Palliative End Points," *Journal of Clinical Oncology*, 14, 1756-1764.

Ten Have, T.R., Miller, M.E., Reboussin, B.A., et al. (2000), "Mixed Effects Logistic Regression Models for Longitudinal Ordinal Functional Response Data with Multiple Cause Drop-Out from the Longitudinal Study of Aging," *Biometrics*, 56, 279-287.

Ten Have, T.R., Pulkstenis, E., Kunselman, A., et al. (1998), "Mixed Effects Logistic Regression Models for Longitudinal Binary Response Data with Informative Drop-out," *Biometrics*, 54, 367-383.

Ware, J.E., Kosinski, M., and Keller, S.D. (1994), *SF-36 Physical and Mental Summary Scales: A User's Manual.* Boston, MA: The Health Institute, New England Medical Center.

201 N. CRAIG STREET, SUITE 350
NSABP BIOSTATISTICAL CENTER
PITTSBURGH, PA 15213
UNITED STATES

Figure Legends:

Fig. 1. Change from baseline score for depression in participants in the BCPT. Depression is slightly increased in the placebo group, compared to the tamoxifen group (not statistically significant).

Fig. 2. The percent of missing questionnaires at each visit by baseline CES-D group (0-10, 11-15, 16-21, and 22-60) which is higher for subjects with higher baseline CES-D scores.

Fig. 3. A. Increase in percent of participants whose CES-D score was at least 16, minus the percent at baseline. The percent increased in both arms.

B. Effect of missing data. Increase in percent of participants whose CES-D score was at least 16, after imputation with the previous observed score, adjusting for the difference in treatment arm means between the missed visit and the preceding visit. The imputed observations in the tamoxifen arm had an additional 3.2% added, and the resulting curves are nearly overlapping.

Fig. 4. The mean hot flash score after subtraction of each participant's baseline score, by treatment arm (solid lines) and the mean hot flash score after subtraction of each participant's baseline score, by treatment arm (dashed lines). Tamoxifen subjects experienced more severe hot flashes. For each subject, missing values were first imputed with previous observed values, adjusting for the difference in treatment arm means between the missed visit and the preceding visit. Imputation did not substantially change the comparison.

Table 3
Number and Percent of Forms Missing by Status of Patient

	Placebo	Tamoxifen
Protocol Event	801 (12%)	791 (11%)
Stopped Therapy	3347 (50%)	3883 (55%)
Withdrew Consent	1359 (20%)	1297 (18%)
On Therapy	1140 (17%)	1134 (16%)
Total	6647	7105

Table 4
Average CES-D Score Prior to Missing vs. Observed Scores

	Avg. CES-D Scores	
	Before Missing Score	Before Observed Score
Tamoxifen Arm	7.78	6.50
Placebo Arm	7.70	6.44

*A score of 16 or higher is considered an indicator of depression.

Table 5

Effect of missing: For CES-D scores, estimates of baseline treatment log odds ratio and corresponding interactions between treatment and time for each follow-up time (standard errors in parentheses) for three models[1]

Model	Base Tx LogOR	Tx by Time Interaction Log OR at Follow-up Times (months)						
		3	6	12	18	24	30	36
Naïve	0.17 (0.08)	-0.07 (0.09)	-0.11 (0.09)	-0.20 (0.09)	-0.18 (0.09)	-0.17 (0.09)	-0.26 (0.10)	-0.19 (0.10)
Joint 1	0.18 (0.08)	-0.08 (0.09)	-0.12 (0.09)	-0.21 (0.09)	-0.19 (0.09)	-0.18 (0.10)	-0.27 (0.10)	-0.20 (0.10)
Joint 2	0.17 (0.08)	-0.08 (0.09)	-0.12 (0.09)	-0.21 (0.09)	-0.19 (0.09)	-0.17 (0.10)	-0.27 (0.10)	-0.19 (0.10)

[1]Models: 1) the random effects logistic model without a drop-out component under the assumption drop-out is MAR (naïve model); 2) the random effects logistic model augmented with a discrete time survival logistic model for drop-out that shares a random effect with the ordinal symptom outcome (joint 1 model); 3) model 2) with the last observed symptom outcome added as a covariate (joint 2 model).

Table 6:

For CESD, estimates of variance components of random intercepts for three models[1]

Model	Symptom Outcome Component	Non-Protocol-Specified Drop-out	Protocol-Specified Drop-out
Naïve	5.25	NA[2]	NA[2]
Joint 1	5.91	0.05	0.01
Joint 2	5.59	0.04	0.02

[1] Models: 1) the random effects logistic model without a drop-out component under the assumption drop-out is MAR (naïve model); 2) the random effects logistic model augmented with a discrete time survival logistic model for drop-out that shares a random effect with the ordinal symptom outcome (joint 1 model); 3) model 2) with the last observed symptom outcome added as a covariate (joint 2 model).

[2] NA: not applicable because naïve model does not include drop-out components

Table 8
Difference in Mean Preceding
Missing HFS – Observed HFS: Logistic approach
would be appropriate if the values all had the same sign

Visit (Mo.)	Placebo	Tamoxifen
3	0.0259	0.0384
6	0.1218	0.1418
12	0.1212	-.0484
18	0.1037	-.0952
24	-.0272	-.0850
30	-0.0594	-.2230
36	.0862	-.1182

Figure 1

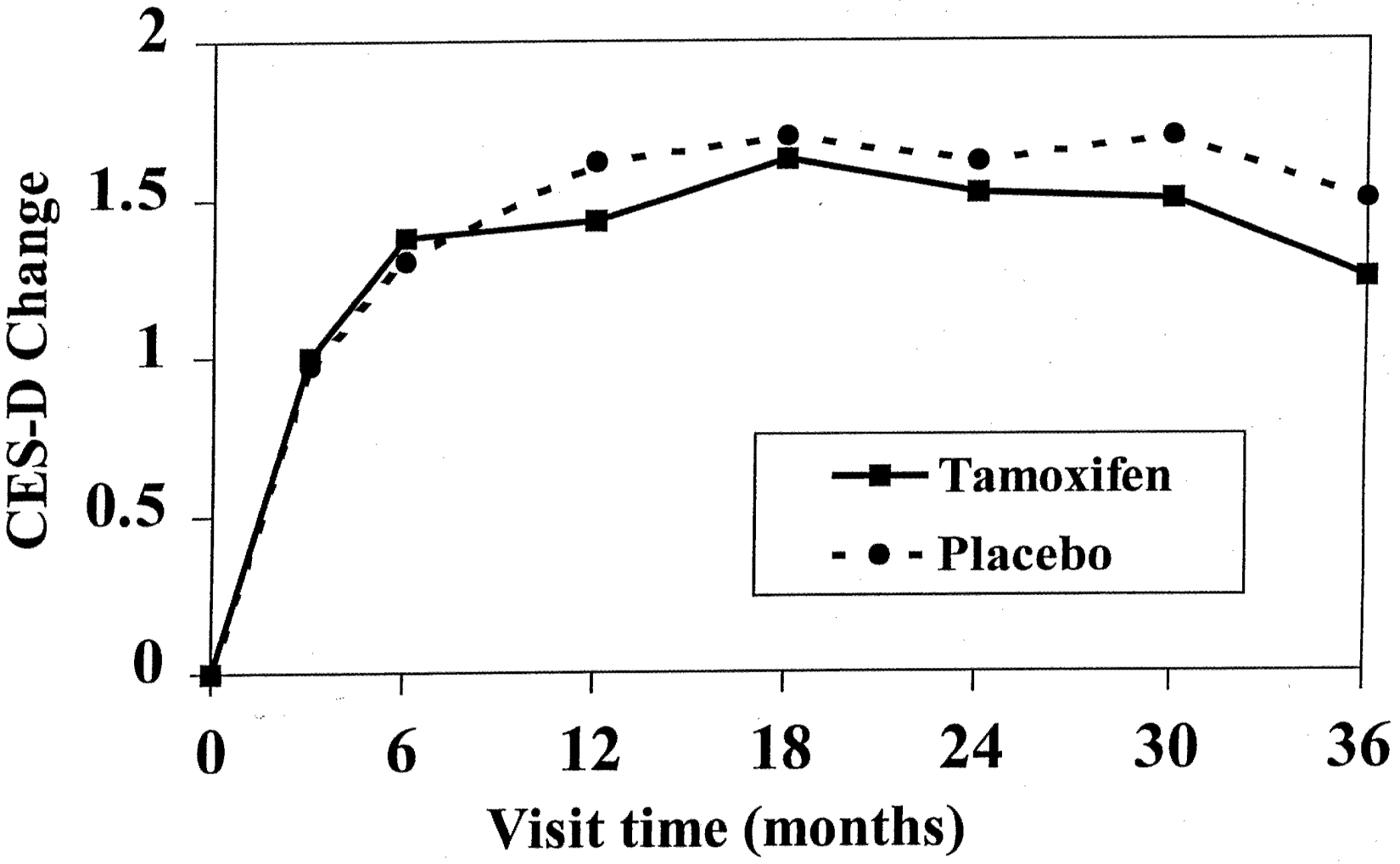

CES-D Change
2
1.5
1
0.5
0
0
6
12
18
24
30
36
Visit time (months)
Tamoxifen
Placebo

Figure 2

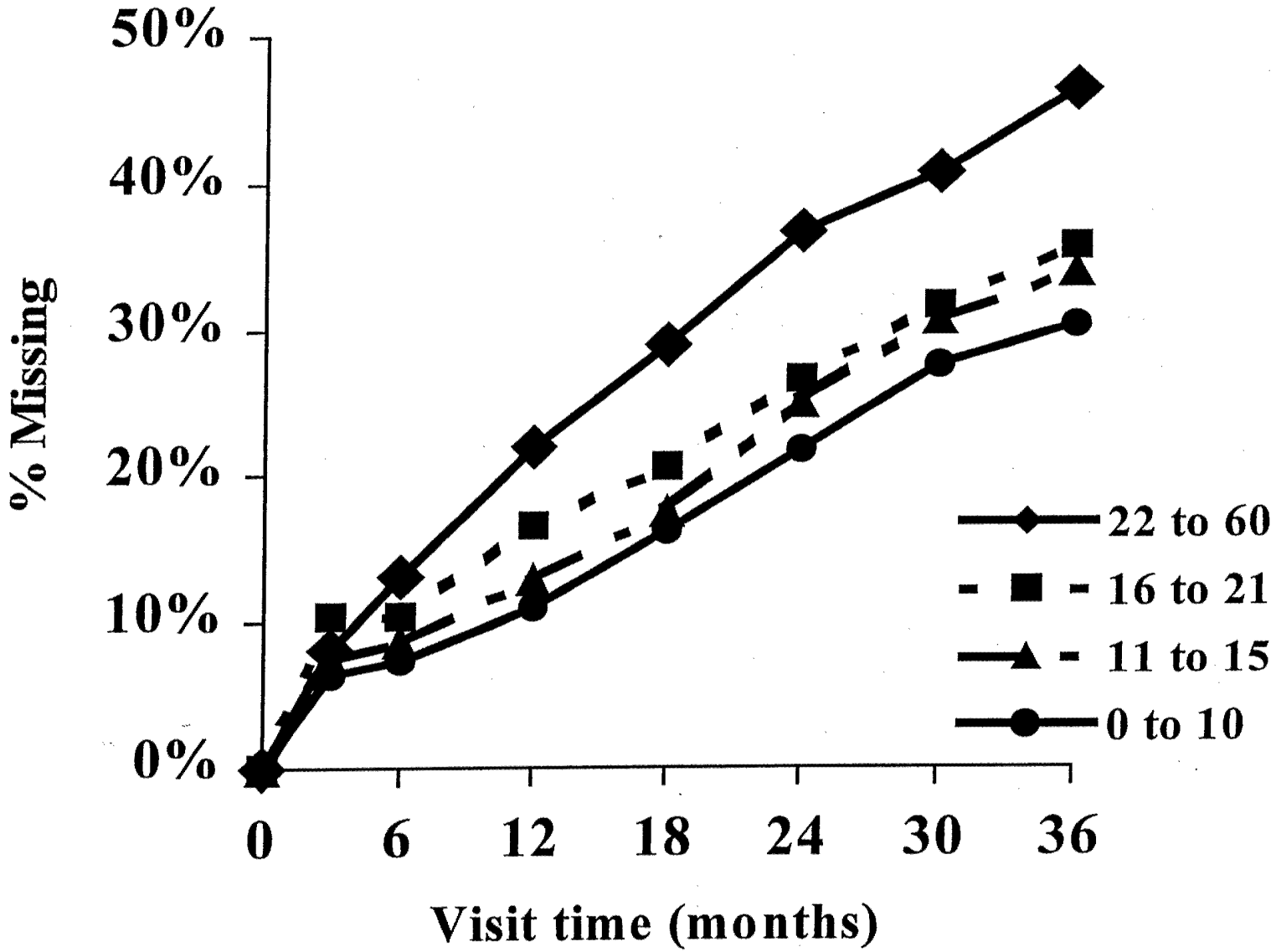

50%
40%
30%
20%
10%
0%
% Missing
0
6
12
18
24
30
36
Visit time (months)
22 to 60
16 to 21
11 to 15
0 to 10

Figure 3.A

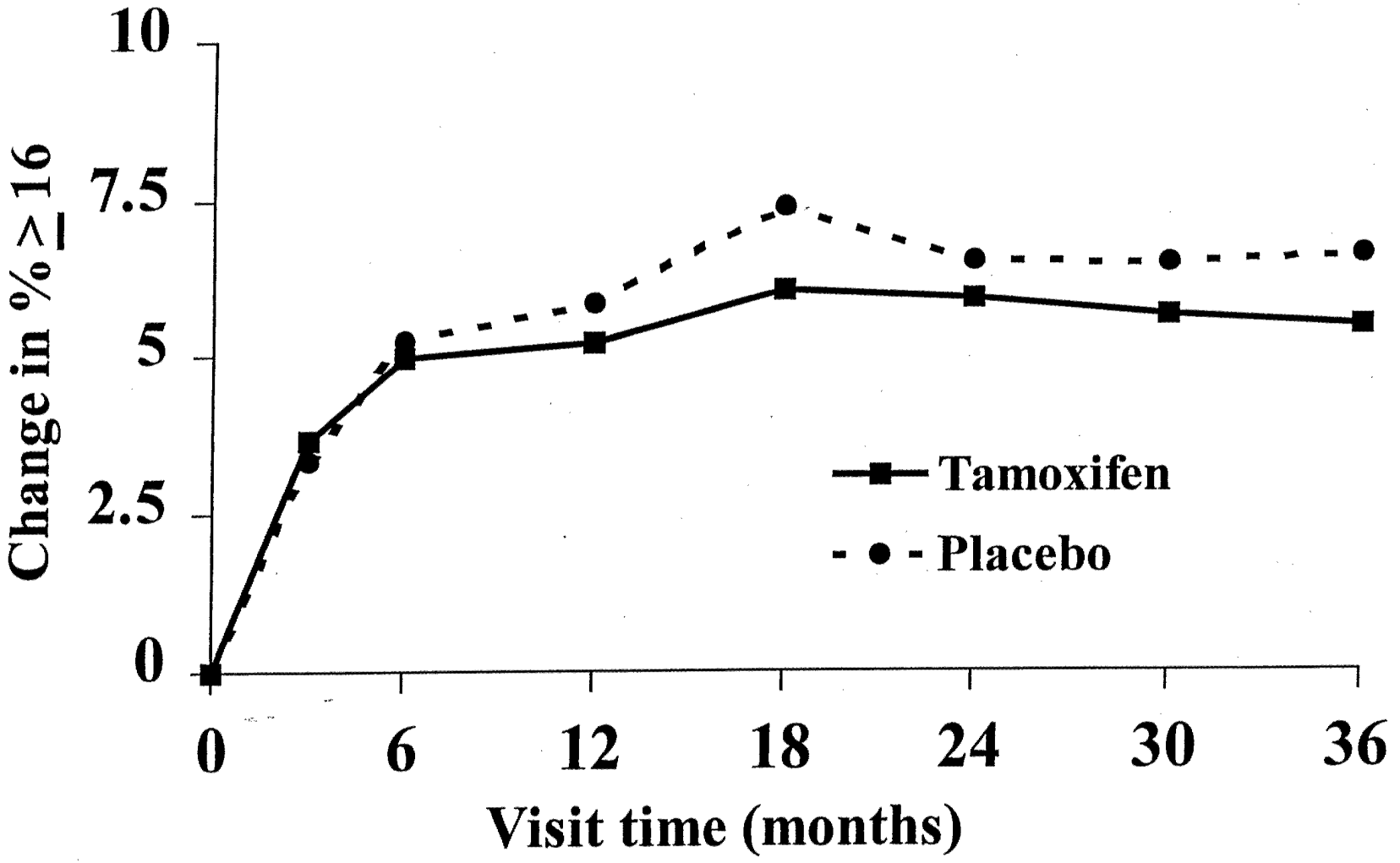

10
7.5
5
2.5
0
Change in % ≥ 16
0
6
12
18
24
30
36
Visit time (months)
Tamoxifen
Placebo

Figure 3.B

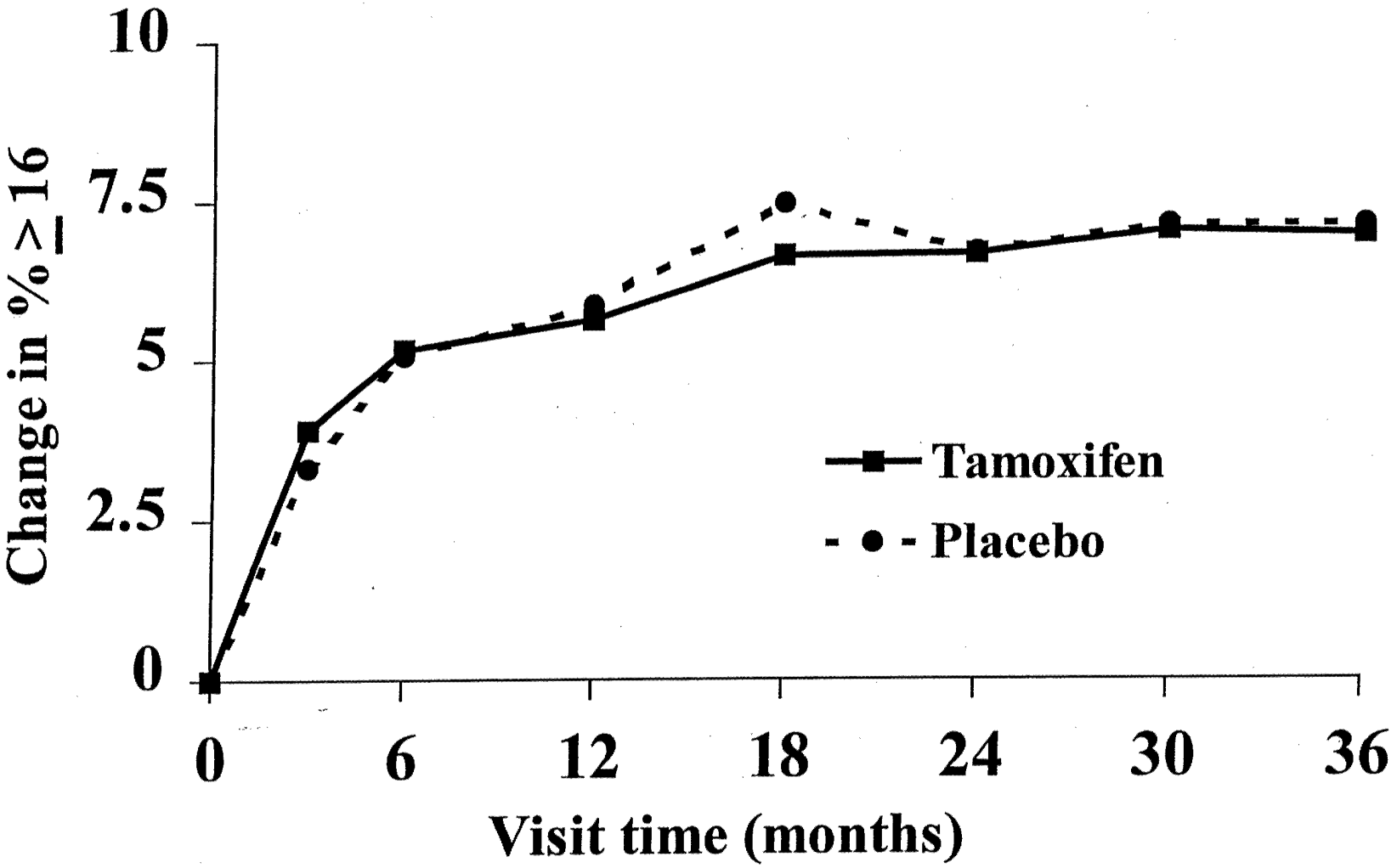
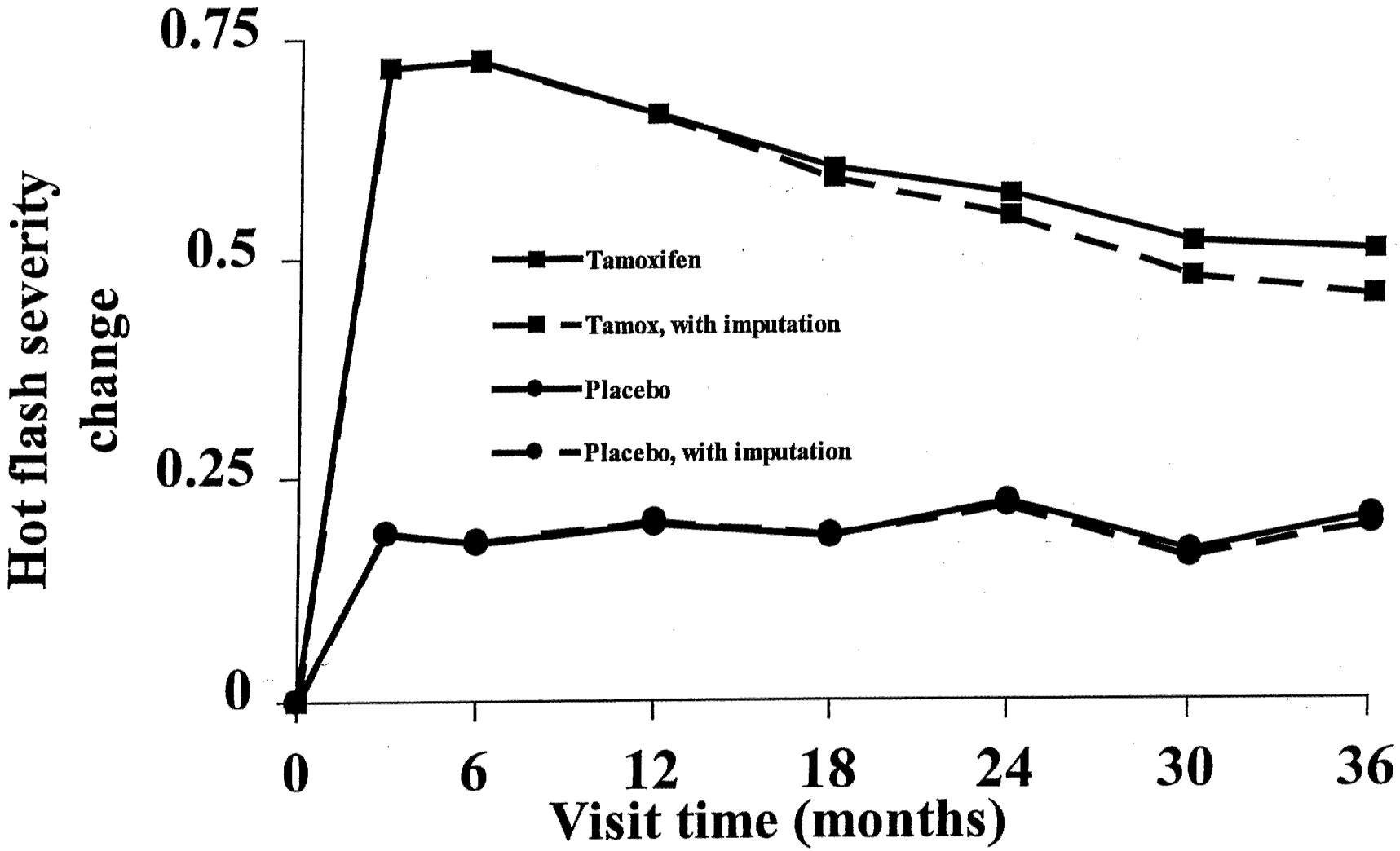

10
7.5
5
2.5
0
Change in % ≥ 16
0
6
12
18
24
30
36
Visit time (months)
Tamoxifen
Placebo

Figure 4


0.75
0.5
0.25
0
Hot flash severity change
Tamoxifen
Tamox, with imputation
Placebo
Placebo, with imputation
0
6
12
18
24
30
36
Visit time (months)

Quality of Life Research 9: 1073–1080, 2001.

Practice and policy of measuring quality of life and health economics in cancer clinical trials: A survey among co-operative trial groups

G. Kiebert[1,2], S. Wait[2], J. Bernhard[3], A. Bezjak[4], D. Cella[5,6], R. Day[7], J. Houghton[8], C. Moinpour[9], C. Scott[10] & R. Stephens[11]

[1]*MEDTAP International, London, UK;* [2]*Quality of life Taskforce in Oncology, Novartis Pharma, Basel, Switzerland;* [3]*Swiss Group for Clinical Cancer Research, and International Breast Cancer Study Group, Bern, Switzerland;* [4]*National Cancer Institute of Canada, Clinical Trials Group, Toronto, Ontario, Canada;* [5]*Eastern Cooperative Oncology Group, Evanston, IL, USA;* [6]*Gynecologic Oncology Group, Evanston, IL, USA;* [7]*National Surgical Adjuvant Breast and Bowel Project, Pittsburgh, PA, USA;* [8]*Cancer Research Council and UCL Cancer Trials Centre, London, UK;* [9]*Southwest Oncology Group, Seattle, WA, USA;* [10]*Radiation Therapy Oncology Group, Philadelphia, PA, USA;* [11]*Medical Research Council, Clinical Trial Unit, London, UK*

Accepted in revised form 2 February 2001

Abstract

Background: Co-operative groups have played an important role in the advance of health-related quality of life (HRQL) research. However, definitions of the concept, criteria for selection of existing instruments and methods for data collection and interpretation remain poorly defined in the literature. A survey was conducted amongst the major cancer co-operative groups in order to gain a better understanding of their current policy and processes to ensure optimal HRQL data collection within cancer clinical trials. The topic of health economics was similarly addressed. *Methods*: A written questionnaire was addressed to 16 major European and North American cancer co-operative groups. Eleven groups responded (response rate: 69%), however, one group could not provide information for the survey, thus ten questionnaires were available for analysis. *Results*: The results from this survey among co-operative groups show that HRQL (more than health economics) is recognized as an important, although usually secondary, outcome measure in oncology trials. On the whole, co-operative groups have a rather flexible policy towards the inclusion of HRQL (and HE) into their clinical trials, and practice is very much on a case-by-case basis, but use standard practice guidelines and internal procedures is to ensure well-defined study protocols and enhance good quality studies.

Key words: Cancer, Co-operative group, Health economics, Randomized controlled clinical trials, Quality of life

Introduction

In chronic diseases where cure is often not achievable, it has long been recognized that improvement in health-related quality of life (HRQL) is of great importance. Oncology was one of the first disease areas where the importance of HRQL as an outcome measure was acknowledged; in US, HRQL outcomes were first included in large treatment and prevention trials in cardiovascular disease. Over the past ten years, there has been an increasing emphasis on the role of alternate outcomes other than the classical clinical trial endpoints of response rate, disease-free or overall survival. Since most trials take many years to mature, it is only now that gradually more and more publications of clinical trials include HRQL.

This study was supported from an unrestricted financial grant from Novartis Pharma AG.

Co-operative groups are playing an increasingly important role in the advancement of cancer care through the conduct of clinical trials, and the establishment of treatment recommendations and guidelines. Collaborative trials groups have also been active proponents of quality of life research. For instance, the proceedings of a workshop focusing on practical and methodological issues related to missing quality of life data in clinical trials in which all major co-operative trial groups participated and contributed were recently published as a special issue in Statistics in Medicine [1].

An informal review of existing literature indicated that many of the large oncology co-operative groups have some kind of policy or guidelines for the inclusion of HRQL as an endpoint in cancer clinical trials. However, the overall information from existing publications is scarce, incomplete and not up-to-date. In particular, information on criteria for selection of existing instruments, methods for assessment, and data collection procedures and instructions is lacking. For this reason, a survey was done of the major co-operative groups (i) that conduct clinical studies in more than one type of cancer or (ii) that focus on a single type of cancer but whose scope and membership are pan-continental.

The objective of this survey was to obtain up-to-date information on the co-operative group policy on HRQL research. Since health economics (HE), specifically resource utilization data collection, is gradually being evaluated in cancer clinical trials, our survey addressed this as an additional topic.

The survey was developed and conducted within the context of a special multidisciplinary taskforce, whose mandate was to develop internal guidelines on HRQL evaluation within oncology clinical trials at a large pharmaceutical company. Recognizing the prominent role that co-operative groups have played in HRQL research in oncology, the taskforce felt that it was essential to look to these groups for 'state of the art' processes and strategies to ensure optimal HRQL data collection within clinical trials.

Methods

The target group consisted of all major national or international co-operative groups that conduct studies in more than one type of cancer and multicontinental groups focusing on one type of cancer. The first step involved the identification of the key person in each co-operative group responsible for quality of life issues who could respond to the questionnaire on behalf of the co-operative group. This step was performed by telephone survey by the principal study investigator (GK). For all groups this key person is a specialized quality of life researcher. Once the key person was identified, this person was sent a cover letter stating the objective and content of the survey, an invitation to participate, and a request to return the completed questionnaire within six weeks. A written reminder was sent to all non-responders after six weeks. Three weeks thereafter, the remaining non-responders were contacted by telephone and, in one case, by fax.

The final response rate was 11 out of 16. Three groups did not respond, two groups refused (one because of time constraints (Cancer and Leukemia Group B (CALGB)) and one because of concerns about confidentiality of information (European Organisation for the Research and Treatment of Cancer (EORTC)). One group was willing to participate, but at the time of the survey this information was not readily available for organizational reasons. Thus, a total of ten questionnaires were available for analysis. Table 1 provides an overview of the groups that were approached and their responses to our invitation to participate in the survey.

The questionnaire was developed especially for this survey. A listing was made of all relevant topics for which we intended to collect data. In a second step a set of questions were formulated addressing all different aspects of each topic. A draft version of the survey was reviewed by members of the taskforce experienced in the development of questionnaires.

The questionnaire addressed the following topics: overview of ongoing clinical trials with and without HRQL in the most prevalent types of cancer; co-operative group trial selection policy; procedures and methods for inclusion of HRQL into clinical trials; study center training and guidelines for HRQL data collection; data analysis and reporting of findings. The same questions were asked for HE. The results of the survey are discussed below in this order of topics.

Results

Overview of ongoing clinical trials

Most numerous of on-going clinical trials are those in gynecological, breast, lung, prostate and colorectal cancers (Table 2). In more than half of these trials, HRQL is evaluated, although usually as a secondary endpoint, and only seldom as the primary endpoint. Notable exceptions are trials that evaluate best supportive care, where HRQL is the primary endpoint in six out of eight trials. HE endpoints are much less frequently collected in the reported trials.

Trial selection policy

Limited research resources and budget constraints often necessitate prioritising of HRQL studies. In the context of clinical trials this situation is not

Table 1. Overview of target groups and survey response

Co-operative group	Response
Cancer Research Council, UCL Cancer Trials Centre (UK)	Yes
Medical Research Council, Clinical Trials Unit (UK)	Yes
Swiss Group for Clinical Cancer Research (SIAKK/SAKK) (Switzerland)	Yes
International Breast Cancer Study Group (Switzerland)	Yes
National Cancer Institute of Canada, Clinical Trials Group (Canada)	Yes
Eastern Cooperative Oncology Group (US)	Yes
Gynecologic Oncology Group (US)	Yes
Southwest Oncology Group (US)	Yes
National Surgical Adjuvant Breast and Bowel Project (US)	Yes
Radiation Therapy Oncology Group (US)	Yes
Deutsche Krebsgesellschaft (as representative of the German Co-operative Groups (Germany))	Willing, but information not readily available
Interdisciplinary Group for Cancer Care Evaluation (Italy)	No response
Fédération Nationale des Centres de Lutte Contre le Cancer (France)	No response
Nordic Cancer Trial Group (Scandinavia)	No response
Cancer and Leukemia Group B (US)	Refused
European Organisation for the Research and Treatment of Cancer (Europe)	Refused

Table 2. Summary of ongoing clinical trials by disease site in 1998 for all ten groups surveyed

Cancer site*	Total number of ongoing trials in 1998 (n)	Number of trials with HRQL as primary endpoint n (%)	Number of trials with HRQL as secondary endpoint n (%)	Number of trials including HE n (%)
Brain	11	0	5 (45)	0
Breast	30	3 (10)	14 (47)	1 (3)
Colorectal	20	1 (5)	9 (45)	3 (15)
Prostate	19	3 (16)	10 (53)	0
Gynecology	32	1 (3)	16 (50)	5 (16)
Head & Neck	16	0	6 (38)	0
Leukemia	10	0	0	0
Lung	27	0	13 (48)	2 (7)
Lymphoma	11	1 (9)	1 (9)	0
Melanoma	3	0	1 (33)	0
Multiple cancer sites				
Supportive care	9	6 (67)	2 (22)	0
Palliative care	9	2 (22)	6 (67)	2 (22)

* Selection of type of cancer based on prevalence of the disease. It does not represent a complete overview of all ongoing clinical trials per group.

different. The trials where HRQL is considered most important are those in which a large survival advantage is not expected; which compare very different treatments (e.g., chemotherapy vs. radiation) that will likely result in different side effect profiles, and those in which patients are symptomatic and the treatment is expected to relieve those symptoms.

In nine out of ten groups, there is a specific person or committee in the co-operative group responsible for HRQL research issues such as trial selection, procedures for data collection, implementation, and methodology. Only one of the co-operative groups has adopted a policy of including HRQL in all cancer clinical trials as a standard (National Cancer Institute of Canada (NCIC)). In all other groups, this decision depends on a number of factors such as study design, research question, sample size, number of participating centres and countries, and a number of population characteristics. A randomized study design allows for comparison of HRQL between the two study arms and distinguishes the effect of trial intervention over time. The research question determines the relevance of HRQL as an endpoint to that question and the sample size distinguishes whether there will be a sufficient number of patients to provide an answer to the HRQL research question. The number of participating centres and countries influence the feasibility of HRQL assessment and likelihood of compliance to questionnaire completion, the number of languages in which the questionnaire will need to be available, as well as funding needed. Duration of the trial affects feasibility as well as funding issues. Financial constraints can play limitating role and necessitate prioritising of trials that include HRQL as an outcome parameter. Age of the patients is most relevant in the paediatric population to determine whether self-assessment of HRQL is possible. And lastly, the health care setting frequently influences the availability of personnel to administer HRQL questionnaires.

Table 3 provides the detailed ratings of importance of different factors in the selection of trials for inclusion of HRQL. Numbers represent the sum of responses from the 10 groups surveyed. Globally, treatment characteristics appear to play a more important role in the selection of trials for HRQL data evaluation than trial and population characteristics. Study design, available resources, toxicity of treatment and absence of incremental survival advantage were the most important factors.

There is often discussion as to whether HRQL is best collected within the actual clinical trial or as a separate or companion protocol. When asked whether HRQL studies were conducted as an integral part of the study protocol, seven groups responded 'yes, always', and three reported 'sometimes'. Six respondents stated that HRQL was never conducted with a separate protocol, and four respondents stated that this was sometimes the case. To the question whether, when included in a trial, HRQL was a mandatory aspect of the study for all participating centers, five groups responded 'yes, always', one 'no, never' and four 'sometimes'.

Mode of assessment and choice of instrument

All but one group use written questionnaires as a standard mode of HRQL assessment, and five groups use in principle the same instrument in all studies (either EORTC QLQ-C30; FACT-G; or LASA scale). For the other groups, the choice of the instrument depends mainly on the trial characteristics, psychometric properties and its practicality for a particular trial, and to a lesser degree on language availability, familiarity with the instrument or its theoretical foundation. Examples of questionnaires that have been used previously in trials are SWOG QoL questionnaire, CARES-SF, MOS-SF36, EORTC QLQ-C30; FACT-G; or LASA scale.

HRQL research guidelines

All groups provide some form of specific instructions to the participating centers for the collection of HRQL data. These can consist of written guidelines, training days, a HRQL training video, procedure manuals for HRQL assessment, regular internal training at group meetings, and an initiation site visit to discuss the HRQL aspects of the protocol.

Six out of ten groups have written internal procedures or guidelines for HRQL data analysis and interpretation. Topics covered by all guidelines include the plan for statistical data analysis

Table 3. Average importance of factors influencing decisions to include HRQL as an endpoint in a clinical trial

	Importance rating			
	Not at all	A bit	Quite a lot	Very much
Trial characteristics				
Resources available	0	2	4	4
Study design	2	2	2	4
Monitoring capacity	2	2	5	1
Representativeness of participating investigators and centers	3	4	1	2
Sample size	1	6	3	0
Participating countries	1	8	0	0
Number of participating countries	2	7	0	0
Number of centers	4	6	0	0
Duration of trial	5	4	1	0
Treatment of characteristics				
Equal efficacy in terms of survival expected	0	0	2	7
Toxicity of treatment	0	0	6	3
New treatment modality	0	1	6	1
New mode of administration	0	5	3	1
Palliative intent	2	1	2	3
Curative intent	2	4	2	0
Population characteristics				
Age (children, adults, elderly)	2	6	1	1
Representativeness of trial population	3	4	3	0
Health care setting (in- vs. outpatient department or home care)	2	8	0	0
Instrument characteristics				
Availability of suitable instrument	0	6	1	3
Other				
Burden on patients				
Statement that HRQL outcome is critical for interpreting results				
Potential outside funding				

Note: Answers shown above represent the sum of respondents choosing that category.

and calculation of sample size estimations. Handling of missing data is included in five out of six. Other topics mentioned were the interpretation of results as clinically meaningful changes over time (n = 1), in relation to clinical data (n = 3) or to other outcome measures (n = 1). Only one group addresses the issue of the pooling of data for multinational analysis, which is not surprising as the majority of respondents are groups that operate mainly at a national level.

Topics that are not addressed at all in existing guidelines are the dissemination of results within clinical practice and the role of HRQL outcomes in subsequent treatment decision making.

Interest in HRQL research

Four groups stated that their interest in HRQL research is very high, and five groups expressed quite some interest (missing n = 1).

Health economics

In general, the activity and interest in health economics is significantly less among all groups than for HRQL. In three groups, health economics data in the form of resource use such as hospitalization, medication, diagnostic tests used, number of outpatient visits, have never been assessed in any trial.

Four groups have a person or committee specifically responsible for health economic issues; one group has a broad outcomes committee that can address health outcomes including HE. None of the groups has a standard policy to collect HE data in each trial.

Four groups identified formal criteria that they followed when deciding whether to include HE as an outcome measure. The most important considerations were the direct cost of the investigated treatment(s), costs associated with treatment of adverse events, and potential financial consequences of treatment for the hospital, practice, or patient. Trial population characteristics and external requirements from health authorities and/or medical ethics committees play a less important role in HE inclusion decisions.

Three groups have some sort of guidelines for the collection of HE data. None of the groups has internal procedures or guidelines for the analysis and interpretation of HE data.

Interest in HE research

The perceived level if interest in HE is fairly low: one group is very interested, two groups are quite interested and five groups indicated a bit of interest in the subject (missing: n = 1).

Discussion

The objective of this study was to obtain up-to-date information of the processes and strategies used by large national and international oncology co-operative groups to conduct HRQL research and to ensure optimal HRQL data collection within their clinical trials. Questions were also asked with regards to the groups' policy towards HE data collection, as it is felt that this is an emerging, and complementary, field of research to that of HRQL [2].

One of the important limitations of our study is the size and representativeness of the study sample. We approached only (i) large national or international co-operative groups that conduct studies on more than one type of cancer and (ii) multi-continental groups focusing on one type of cancer. Moreover, we did not include groups active in the field of pediatric oncology. As a result, there are clear limitations regarding the representativeness of our sample and the generalizability of the results. The majority of the participating co-operative groups is North American, leaving other continents, and especially Europe, clearly underrepresented. Non-participation in our survey does not imply lack of experience or policies regarding HRQL and HE research. For instance, the EORTC has been active in the field of HRQL research since many years, and has published on their strategy to include HRQL as an endpoint in their clinical trials [3]. It would be inappropriate to infer their policy from publicly available information as these will not provide the same level of detail obtained by our survey. The same approach would also have to be applied to other co-operative groups, and published reports from other multinational or national European groups on HRQL and HE policies and strategies are scarcer.

One may ask the question whether Europe is different from North America in its approach to HRQL research. One source of information is to look at the stance of health authorities to HRQL in these two continents. In US, a 1996 publication [4] on the position of the Federal Drug Administration (FDA) with regards to HRQL suggests that, for the FDA, HRQL is more important than traditional measures of efficacy such as tumor response for drugs that do not have any impact on survival. More recently, the FDA has set up a special committee in collaboration with outside researchers to investigate further the role of HRQL within the registration and labeling of oncology products (i.e. Subcommittee of the Oncology Drug Advisory Committee). In Europe, the European Medicines Evaluation Agency (EMEA) cite "symptom control backed up by quality of life assessments" as one of the possible secondary outcome measures in their 1996 publication of the Committee for Proprietary Medicinal Products (CPMP) [5]. However, the actual role that HRQL data have played in drug approval decisions by both of these agencies remains to be elucidated [6]. One positive example in the US is the role played by HRQL data, specifically reduction of pain, in the approval of mitoxantrone and prednisone for the treatment of hormone-refractory prostate cancer [7]. Indeed, it may be assumed that authorities in both Europe and North America are at the early stages of learning about the value of

HRQL research and findings to the development and acceptance of new cancer therapies. Within this learning environment, co-operative groups in all continents may play an important role in setting precedents, disseminating research findings and advancing methodologies in this growing field.

In our survey, we did not ask respondents to differentiate between trials that are financially supported publicly or by the pharmaceutical industry. Clinical trials in US are predominantly sponsored by the government, whereas co-operative groups in Europe and Canada have more of a mixture of government and industry sponsored studies. For industry sponsored trials, the most influential factor on whether to include HRQL as an endpoint is the requirement of this type of data by the regulatory authorities. From the perspective of the co-operative group, the issue of available funding is of great importance and can to a certain extent influence the support for HRQL assessments. Industry reimbursement rates per patient participating in a trial are usually greater than funding rates from public sources and the added resources can be used to pursue non-traditional endpoints or to provide financial support for studies involving non-pharmaceutical therapeutic modalities. It would be very interesting to conduct a similar survey among pharmaceutical companies and to be able to compare the pharmaceutical policies regarding the inclusion of HRQL and health economic research questions in clinical trials to those of co-operative groups.

The results from this survey among co-operative groups show that HRQL is recognized as an important, although usually secondary, outcome measure in oncology trials. Although health economics data such as hospitalizations or other resource use play a much lesser role in the clinical trial context, their role in reimbursement decisions may be more prominent. On the whole, co-operative groups have a rather flexible policy towards the inclusion of HRQL (and HE) into their clinical trials, and practice is very much on a case-by-case basis. The fact that many groups have developed written internal procedures or guidelines does not mean that they adopt a rigid approach towards design, analysis or interpretation of results. The purpose of guidelines and internal procedures is to ensure well-defined study protocols and enhance good quality studies. This is underlined by the fact that all groups recognized the importance of training of clinical trial managers for HRQL data collection, an aspect often neglected in industry-run HRQL studies. The fact that HRQL evaluation was most often recognized as an integral, and often mandatory, part of clinical trials is a promising sign, as acceptance and understanding of this outcome by treating physicians will only grow with their increased exposure to its analysis within the context of other clinical findings.

One aspect that was not addressed by all groups was the dissemination and positioning of HRQL findings within the context of clinical trial evidence and the implications of these findings for clinical practice. The need for further research and guidance in this area was also highlighted in several surveys of practicing oncologists on their perception of HRQL [8, 9]. Clearly, an essential aspect to the development of HRQL research remains the proper interpretation of findings, clear communication of the results to practicing physicians and patients, and, ultimately, the integration of HRQL aspects of therapy into actual treatment decisions.

In conclusion, the results of this survey confirm the impression that HRQL research is a growing, however still developing field in the context of clinical trials. Co-operative groups are likely to continue to play an increasing role in the advancement of this science and the dissemination of findings to treating physicians and their patients. Their role in the promotion of health economics research may be a lesser one. One may hope that the knowledge and experience that these trials groups acquire in including HRQL parameters into their trials may serve other researchers and drug sponsors in achieving a more comprehensive assessment of the impact of new therapies on cancer patients.

References

1. Statistics in Medicine 1998: 17.
2. Fitzpatrick R, Davies L. Health economics and quality of life in cancer trials: Report based on a UKCCR workshop. Br J Cancer 1998; 77(10): 1543–1548.
3. Kiebert GM, Kaasa S. Quality of life in clinical cancer trials: Experience and perspective of the European organization for research and treatment of cancer. J Nat Cancer Inst Monographs 1996: 20: 91–97.

4. Beitz J, Gnecco C, Justice R. Quality of life end points in cancer clinical trials: The US food and drug administration perspective. J Nat Cancer Inst Monographs 1996; 20: 7–9.
5. EMEA. Recommendations for Efficacy Measurement in Oncology 1995.
6. Moinpour CM, Ganz P, Rowland J, Gritz E, Gotay C. Guest editorial: The value of quality of life data in judging patient benefit: Experts respond to ODAC. The Cancer Letter 1999; 25: 1–5.
7. Tannock IF, Osba D, Stockler MR, et al. Chemotherapy with mitoxantrone plus prednisone or prednisone alone for symptomatic hormone-resistant prostate cancer: A randomized trial with palliative endpoints. J Clin Oncol 1996; 14: 1756–1764.
8. Taylor KM, Macdonald KG, Bezjak A, Ng P, DePetrillo AD. Physicians' perspective on quality of life: An exploratory study of oncologists. Qual Life Res 1995; 5: 5–14.
9. Morris J, Perez D, McNoe B. The use of quality of life data in clinical practice. Qual Life Res 1998; 7: 89–91.

Address for correspondence: G. Kiebert, MEDTAP International, 27 Gilbert Street, London W1Y 1RL, UK
Phone: +44 20 7290 9405; Fax: +44 20 7290 9705
E-mail: kiebert@medtap.co.uk

B-23 QOL study 4:11 PM 12/9/02

NOT A FINAL DRAFT.

Health-Related Quality of Life in Axillary Node-Negative, Estrogen Receptor-Negative Breast Cancer Patients Undergoing AC versus CMF Chemotherapy: Findings from the National Adjuvant Breast and Bowel Project B-23

Tentative author list: SR Land, J Kopec, G Yothers, S Anderson, R Day, G Tang, P Ganz, B Fisher, N Wolmark

This investigation was supported by Public Health Service grants ?? from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services.

Address correspondence and reprint requests to Stephanie R. Land, NSABP, 201 N. Craig St., Pittsburgh, PA, 15213. Phone: 412-383-2213; fax: 412-383-1387; e-mail: land@nsabp.pitt.edu

ABSTRACT

Purpose: NSABP Protocol B-23 compared two chemotherapy regimens: (1) cyclophosphamide, methotrexate, and 5-fluorouracil (CMF) and (2) doxorubicin (Adriamycin) and cyclophosphamide (AC) in terms of relapse-free survival, event-free survival and overall survival. The AC regimen is shorter in duration and was expected to be less toxic than the CMF regimen. Patients in B-23 were node-negative and estrogen receptor-negative. There is no previous data regarding the trade-off in quality of life between the two regimens in this population of breast cancer patients. Quality of life information was considered especially relevant given the possibility that the two chemotherapy regimens would prove equivalent in terms of clinical outcome.

Patients and Methods: One hundred sixty patients participated in the NSABP B-23 Quality of Life study. Patients in B-23 were randomly assigned to CMF plus tamoxifen (TAM), CMF plus tamoxifen placebo, AC plus TAM or AC plus tamoxifen placebo. Six cycles of CMF were given for 6 months; four cycles of AC were administered for 63 days. TAM or tamoxifen placebo was given daily for 5 years. The Quality of Life questionnaire was administered through the first year after entry to Protocol B-23. The questionnaires included the Functional Assessment of Cancer Therapy (FACT-B), the vitality scale from the Medical Outcomes Study 36-item Short Form Health Status Survey (MOS SF-36), a symptom checklist, and additional items regarding overall quality of life, general health, physical limitations and return to normal activity. Statistical comparisons between treatment arms were performed with area under the curve analyses, repeated measures analyses, and Fisher exact tests.

Results: Bladder problems and diarrhea were significantly more common among patients treated with CMF. In addition, the pattern of quality of life scores over time during treatment differed between chemotherapy treatment groups, with the CMF group having more constant values while the AC group quality of life scores dropped during treatment and rose after treatment. This was especially true with respect to energy levels, which were up to 10 points lower (on a scale of 0-100) in the AC arm. Otherwise, no differences were found between AC and CMF for any of the individual symptoms or any of the other quality of life instruments in terms of (1) overall quality of life during the first nine months after randomization, (2) the average quality of life during treatment, or (3) the rate of recovery to baseline levels of quality of life one year after randomization.

Conclusion: Overall quality of life is equivalent between the two chemotherapy regimens, with a trade-off between more frequent gastrointestinal symptoms in the CMF arm and greater variability in fatigue in the AC arm.

Keywords: Tamoxifen, FACT-B, SF-36

INTRODUCTION

This is the report of the health-related quality of life (HRQOL) component of the National Surgical Breast and Bowel Project (NSABP) Protocol B-23. Protocol B-23 compared the efficacy of four courses of doxorubicin (Adriamycin) and cyclophosphamide (AC) with six courses of standard cyclophosphamide, methotrexate, and 5-fluorouracil (CMF) in women with node-negative, estrogen-receptor (ER)-negative breast cancer. In addition, Protocol B-23

compared the efficacy of tamoxifen (TAM) versus placebo in each chemotherapy group. In addition to chemotherapy, lumpectomy patients received radiation therapy to the breast.

The primary objectives of Protocol B-23 were to compare relapse-free survival, event-free survival and overall survival between chemotherapy regimens, and between tamoxifen and placebo. A detailed description of patient eligibility requirements, study design, therapy used, follow-up, study end points, and statistical analyses of the primary study endpoints appear in Fisher et al, Journal of Clinical Oncol., Feb 2001. The conclusions drawn from the primary analyses were that there was no significant difference in the outcome of patients who received AC or CMF; and TAM with either regimen resulted in no significant advantage over that achieved from chemotherapy alone. The chemotherapy treatment of choice might then depend on other characteristics of the regimens, such as toxicity and quality of life. The AC regimen had been expected to be more tolerable than the CMF regimen due to its decreased duration and toxicity. Measures of toxicity available from the parent B-23 study appear to validate this expectation. More women who began CMF discontinued chemotherapy (10%) than did women who received AC (4%). Approximately 10% more patients in the CMF arm than on the AC arm of Protocol B-23 experienced grades 3 and 4 toxicity.

To complement institution-reported measures of tolerability, a self-administered quality of life questionnaire was administered to a subset of patients enrolled in Protocol B-23. This report provides a comparison of the treatments in terms of self-reported symptoms and physical, emotional and social functioning. The primary comparison of interest is between the two chemotherapy regimens. The effects of other covariates and tamoxifen are of secondary interest, particularly since tamoxifen did not prove to be a clinically efficacious treatment in this patient population.

PATIENTS AND METHODS

Trial design

Women at participating NSABP institutions in the United States and Canada who had primary operable, histologically node-negative, ER-negative breast cancer were enrolled in the parent B-23 study between May 12, 1991 and December 31, 1998. Patients were randomly assigned to one of four treatment groups after surgery: six courses of CMF plus placebo, six courses of CMF plus TAM, four courses of AC plus placebo, or four courses of AC plus TAM. A total of 2,008 patients were randomized to the B-23 Protocol (1,005 to CMF and 1,003 to AC). The treatment assignments were balanced with respect to age (≤ 49 or ≥ 50 years), clinical tumor size (< 2 or ≥ 2 cm), and type of surgery (total mastectomy and axillary node dissection, or lumpectomy and axillary node dissection followed by breast irradiation). Patients who were treated with lumpectomy also received breast irradiation. When administered in conjunction with CMF, radiation therapy was begun within 1 week after day 8 of the first cycle of CMF when there was no evidence of hematologic toxicity. Subsequent doses of CMF were delayed until radiation therapy was completed and until evidence of hematologic toxicity was absent. Breast irradiation was begun after the completion of all AC therapy when blood counts permitted. In all treatment groups, the administration of TAM or placebo was continued during radiation.

The quality of life sub-study was opened in May of 1997. The target accrual for the quality of life evaluation was 200 evaluable patients. However, the B-23 parent study was closed early because it had become apparent that tamoxifen was not effective in this ER-negative population, so accrual totalled only 169 when the protocol was closed on December 31, 1998. Patient and treatment assignment information for the quality-of-life study is listed in Table 1. Slightly more than one half of the patients in the B-23 quality of life study were randomly assigned to AC. Of the AC patients, about one-half were assigned to TAM. Among CMF patients, more than half were randomly assigned to TAM. Of the 169 patients enrolled in the study, 4 were ineligible. There were no submitted questionnaires for 9 enrolled patients.

Table 1. Quality of Life Study Information

	CMF			AC		
Patients	Total number of patients	Placebo number of patients	TAM number of patients	Total number of patients	Placebo number of patients	TAM number of patients
Consented and randomized	82	38	44	87	44	43
Eligible	81	38	43	84	43	41
With at least one questionnaire completed	79	35	44	81	42	39

For the 160 patients with at least one questionnaire submitted, the distributions of the stratification variables are shown in Table 2 according to chemotherapy arm. Age, clinical tumor size and selection of primary surgery are balanced across treatment arm, as would be expected.

Table 2. Distribution of stratification variables according to chemotherapy arm

	CMF	AC
Mean Age (range)	48 (25-73)	49 (28-72)
Mean clinical tumor size (range)	1.9 cm (0-5)	2.0 cm (0-4.5)

	CMF	AC
Lumpectomy, # patients	46	47
Modified radical mastectomy, # patients	33	34

Quality of life assessment schedule

The schedule of chemotherapy and quality of life assessment is shown in Table XX. Quality of life was assessed at the beginning of each chemotherapy cycle and at several follow-up time points. This assessment schedule was designed to allow comparisons of quality of life during chemotherapy treatment and to investigate the longer term differences in quality of life.

Instruments

The quality of life evaluation included the Functional Assessment of Cancer Therapy (FACT-B, Cella, 1993), a symptom checklist, the vitality scale and items regarding general health from the Medical Outcomes Study 36-item Short Form Health Status Survey[2] (MOS SF-36), and additional items regarding overall health-related quality of life and return to normal activity. Each instrument will now be described in greater detail.

1) The FACT-B questionnaire has 30 items divided into five subscales for physical well-being, social/family well-being, relationship with the physician, emotional well-being, and functional well-being. Nine additional questions refer to problems commonly experienced by women with breast cancer. The scale has been extensively validated.[3] Higher values indicate better quality of life. 2) The 17 items on the symptom checklist were selected from several existing instruments, including the symptom checklist from the NSABP Breast Cancer Prevention Trial. The selected symptoms are those commonly reported by patients with cancer, especially patients undergoing surgery and radiotherapy for breast cancer, standard chemotherapy, and hormonal therapy. Each symptom checklist item is scored on a 5-point Likert scale, where higher values indicate more severe symptoms. 3) The four-item SF-36 vitality scale is useful in detecting common side of effects of cancer therapy, such as fatigue and lack of energy. Each item is scored on a 5-point scale, and the sum is rescaled to a range of 0-100, with high values indicating greater vitality. 4) The overall health-related quality of life (HRQOL) item is scored on an 11-point linear rating scale anchored at "death" (0) and "perfect health" (11). 5) The item regarding return to normal activity (developed by D. Cella, personal communication) is also scored on an 11-point scale, with high values indicating greater resumption of normal activities. 6) The MOS SF-36 general health item ("In general, would you say that your health is…") and 7) the MOS SF-36 comparative health item ("Compared to 6 months ago, how would you rate your health in general now?") are each scored on a 5-point scale, with higher values indicating worse health.

There were two versions of the B-23 quality of life questionnaire. At baseline and after the completion of all courses of chemotherapy, the FACT-B, SF-36 vitality scale and symptom checklist items referred to the patients' experience during the prior 7 days, while the overall HRQOL linear rating scale referred only to the day of questionnaire completion. During chemotherapy treatment, the FACT-B, SF-36 vitality scale, symptom checklist and HRQOL linear rating scale referred to the patients' experience since the beginning of the previous chemotherapy cycle. In addition, the comparative health question was not included on the questionnaire given during chemotherapy.

Statistical methods

Analyses included data from all randomized patients who consented to participate in the quality of life study and who completed any quality of life assessments. Due to missing data, some analyses were performed with subsets of these data, as indicated below in the description of each analysis. In addition, questionnaires that were completed after a breast cancer recurrence or second primary cancer were not included in these analyses.

For longitudinal analyses, "time" was defined with respect to chemotherapy treatment. For example, the "week 4" assessment for a patient on the CMF arm referred to the assessment on day 1 of cycle 2, regardless of when that

---

[2] Ware 1994

[3] Provide reference.

assessment was actually performed. This was true even for the subset of patients whose chemotherapy was delayed due to radiation.

1. Missing data

Item non-response was infrequent in this study. Six of 900 questionnaires were missing some part (more than half of any subscale) of the FACT-B, and just 3 of 900 were missing more than half of the SF-36 vitality scale. For the FACT-B subscales and symptom checklist, when an individual item in a scale was missing, the average of other items in the scale was imputed. The SF-36 vitality scale was considered missing if more than two of the four items were missing.

Missing questionnaires were a more frequent occurrence. The statistical analyses described in this report rely on the assumption that the missing questionnaires were missing at random, that is, that patients did not miss an assessment for reasons related to their quality of life. This assumption was tested in several ways. First, the proportion of missing data over time was compared graphically between treatment groups. Second, a t-tests were performed to compare quality of life scores preceding a missed assessment and quality of life scores preceding a completed assessment. For each item or scale considered the t-tests were performed separately at each time point and within chemotherapy groups. Third, a statistical test (Little 1988) was performed for several of the instruments and items to determine whether the missing scores could be considered missing completely at random (MCAR, Little, 1987). This was performed for each chemotherapy group separately, since patients in each chemotherapy group followed a different assessment schedule. Due to the small sample size, only the first four time points were included in the analysis.

2. Overall quality of life: Area under the curve (AUC) analysis

The overall comparison of quality of life between the chemotherapy treatment arms was performed with an area under the curve (AUC) analysis. For a given quality of life scale, this was accomplished in two stages. The first stage is the imputation of missing values. The second stage is the computation and comparison of the areas under the quality of life curves.

The imputation of missing values was performed with repeated measures modeling. Models were estimated for each treatment group separately, including the week of assessment as a factor. This modeling was performed using all available data and all assessment time points. Type of surgery, tamoxifen treatment arm and age were included as predictors in the models. Imputation was then performed by replacing missing values with patient-specific predictions from the models.

The area under the curves was computed for patients whose assessments were performed at baseline, at least once during treatment, and at least once subsequent to treatment. There were 57 such patients in the AC arm and 45 in the CMF arm. Patients with fewer assessments were not included in the second stage of this analysis. In addition, the area was computed only up to the 9-month assessment time point because the rate of missing data was substantial at one year. The AUC was corrected for the baseline score so that a negative AUC value indicates that a patient's scores were lower on average than her baseline score.

Finally, the area under the curve was compared between treatment arm using analysis of variance (ANOVA) with factors for chemotherapy arm; clinical tumor size; age; tamoxifen arm; type of surgery; the interaction between chemotherapy arm and tamoxifen arm; and the interaction between chemotherapy arm and type of surgery.

3. Quality of life during treatment

In order to compare the time course and average levels of quality of life between the two chemotherapy arms during treatment, repeated measures analyses were performed.

The first set of repeated measures analyses were performed to 1) identify the best statistical model for the relationship between quality of life scores and time during weeks 3-20 and 2) determine whether the shape of the quality of life curves were different between treatment arms during weeks 3-20. The effect of time was modeled with a quadratic or linear polynomial. (For each quality of life scale, the best-fitting model for the time effect was

selected from among polynomials and regression spline models[4]). The fixed effects in the model were chemotherapy arm, age, baseline quality of life, tamoxifen treatment arm, tumor size (less than or equal to 2 cm versus greater than 2 cm) and type of surgery (mastectomy or lumpectomy) as well as the interactions between chemotherapy treatment arm and time and between chemotherapy arm and type of surgery. The significance of the fixed effects was tested using maximum likelihood methods.

Repeated measures modeling was also used to compare the overall level of during treatment only (weeks 4-20 for CMF patients, weeks 3-9 for AC patients). The fixed effects in the model were chemotherapy arm, age (at most 50 versus over 50 years old), baseline quality of life, tamoxifen treatment arm, tumor size (less than or equal to 2 cm versus greater than 2 cm) and type of surgery (mastectomy or lumpectomy). The significance of the fixed effects was tested using maximum likelihood methods.

4. Recovery to baseline

The rate of recovery to baseline one year after randomization was compared between the two chemotherapy treatment arms. Recovery to baseline is defined as having a quality of life score that is at least as favorable as the score at baseline. Patients were classified by treatment arm and recovery to baseline. The comparison between treatment arms was then performed with Fisher exact tests.

5. Individual symptoms

The scores for each of the 17 symptoms from the symptom checklist and seven symptom-related items from the FACT-B, was dichotomized at each assessment time point by whether the severity was reported as "not at all". Symptoms were each then compared across chemotherapy treatment arm in a logistic repeated measures model, controlling for the baseline presence of the symptom, tamoxifen treatment arm, age (greater or less than age 50), type of primary surgery (mastectomy versus lumpectomy) and clinical tumor size (less than or equal to 2 cm versus greater than 2 cm). P-values were compared against a Bonferroni-corrected significance level of 0.002 (0.05/24).

RESULTS

Missing data

There was no statistical evidence that missing data would bias the comparisons across chemotherapy treatment arm. Figure XX displays the percentage of completed questionnaires (of those expected) at each assessment time point for each chemotherapy treatment group. The rate of missing questionnaires is approximately equivalent between the chemotherapy treatment groups.

There was no difference in the mean of quality of life items (for any of the seven instruments and additional items in the questionnaire) preceding a missed assessment and the mean of quality of life items preceding a completed assessment. Of the 95 comparisons (one for each instrument, at each relevant time point and for each chemotherapy group), three were significant at the 0.05 level (fewer than would be expected by chance alone), and none was significant at 0.05/95, the Bonferroni-corrected significance level.

The tests of Little (1988) were applied to the FACT-B, symptom checklist total score, SF-36 vitality scale, overall HRQOL linear rating scale, and resumption of normal activities item. These tests revealed no statistically significant evidence against the assumption that missing questionnaires are missing completely at random (all p-values were greater than 0.1).

---

[4] refc

Overall quality of life

There was no difference between chemotherapy treatments or tamoxifen versus placebo in terms of the area under the quality of life curves for the first nine months of the study for any of the following instruments or individual items: FACT-B, SF-36 vitality scale, symptom checklist (average of non-missing items), overall HRQOL linear rating scale and resumption of normal activities scale. There was also no difference between tamoxifen and placebo, and no significant effect of stratification variables on these endpoints.

Quality of life during treatment

Quality of life "during treatment" analyses were performed in two ways. The first analyses include the time period in which at least some patients were on therapy (weeks 3-20), to compare the experience of patients during a comparable time frame. This would potentially allow any improvements in quality of life after the completion of AC to balance more severe quality of life decrements during AC treatment. Both the time course of quality of life and the average level were compared during the week 3-20 time frame. The second comparison included only assessments performed during therapy for each treatment arm (weeks 4-20 for CMF patients, 3-9 for AC patients). As discussed above, "week" reflects the course of therapy at the time of the assessment rather than calendar time, so that "week 20" refers to the assessment that was performed on day 1 of cycle 6 for a CMF patient, regardless of whether there had been treatment delays. All analyses were performed for the FACT-B scale, the SF-36 vitality scale, the symptom checklist scale, the return to normal activity scale, and the overall HRQOL linear rating scale.

Boxplots of the FACT-B scores of patients in each chemotherapy arm at assessment points during weeks 3-20 are shown in Figure WW. The chemotherapy arm did not affect either the average level of the FACT-B score or the shape of the time course of the FACT-B during the weeks 3-20, nor the average level of the assessments during treatment (weeks 4-20 for CMF, 3-9 for AC).

Figure SSS shows the boxplots of the SF-36 vitality scores of patients in each chemotherapy arm at assessment points during weeks 3-20. The time course of the SF-36 vitality scores were significantly different during weeks 3-20 ($p<0.0001$). The largest difference between treatment arms is during weeks 5-10, with median CMF scores up to 10 points higher (on a scale of 0-100, where higher scores indicate greater vitality). The average level of the SF-36 vitality score was not significantly different between treatment arms during weeks 3-20, nor was it significantly different during treatment.

The time course of the symptom checklist scale was also significantly different (data not shown; $p<0.0001$), with symptoms remaining more constant in the CMF arm while they rose during weeks 6-9 and subsequently fell for patients in the AC arm. However, the differences between median scores were very small (approximately 0.1 on a scale of 0-4) throughout weeks 3-20. In addition, the average level of symptoms was not different between AC and CMF patients either during weeks 3-20 or during treatment only.

There were no significant differences between chemotherapy treatment arms in terms of the time course of the "return to normal activities" item, its average level during weeks 3-20, or its average level during treatment (data not shown).

The time course of the health-related linear rating scale was marginally significantly different between chemotherapy treatment arms (data not shown; p=0.046) with scores remaining more constant in the CMF arm while they fell during weeks 6-9 (indicating decreasing quality of life) and subsequently rose for patients in the AC arm. The differences between mean scores were moderate throughout weeks 3-20, with the AC arm at most 0.4 points higher on a scale of 0-10. In addition, the average level of the scores was not different between AC and CMF patients either during weeks 3-20 or during treatment only.

The baseline quality of life was a highly significant predictor of the subsequent quality of life scores for all the quality of life endpoints tested (FACT-B, SF-36 vitality scale, symptom checklist, return to normal activity, and HRQOL linear rating scale). Quality of life significantly diminished during treatment in both chemotherapy treatment groups ($p<0.0001$) for all endpoints except the resumption of normal activity scale (p=0.3). Tamoxifen arm, surgery (mastectomy versus lumpectomy), tumor size and age (whether included as a continuous covariate or as a binary factor, at most 50 versus over 50 years old) did not significantly affect any of the quality of life scales during weeks 3-20 or during treatment only.

Recovery to baseline

The comparisons of the proportion of participants who recovered baseline levels of quality of life (for all seven instruments and items) are shown in Table YY. There is a slight suggestion that more patients on CMF recovered on the FACT-B scale and on the overall HRQOL linear rating scale at one year after randomization, but there is no significant difference between treatment arms.

Symptoms

For each symptom, the odds ratios for chemotherapy are shown in Figure ZZ. Based on a Bonferroni-corrected significance level criterion (0.002), only diarrhea and bladder problems were significantly different between treatment arms. All were greater in the CMF arm, with odds ratios of 3.8 for diarrhea (p-value 0.0001) and 4.2 for bladder problems (p-value 0.0002). There was also a significant decrease in diarrhea incidence among tamoxifen patients as compared to placebo (odds ratio 0.32, p-value 0.0007). The stratification variables did not significantly affect the symptoms.

DISCUSSION

The major differences between chemotherapy treatment groups were that AC patients experienced a greater increase in fatigue and loss of energy during chemotherapy, and a steeper recovery after chemotherapy, while CMF patients experienced more diarrhea and bladder problems. These results were consistent with expectations based on previous experience with the

agents. Severe diarrhea (grades 3-4) was also monitored by the institutions in B-23. They reported rates about 1% higher in the CMF treatment group, which is consistent with --although much smaller than-- the difference seen in the Quality of Life study (odds ratio of 3.8 for being bothered at least "a little bit").

The Quality of Life study did not find a significant difference in self-reported nausea between treatment arms. While CMF was expected to cause more nausea, the results from the parent B-32 protocol suggest that nausea was also better controlled by medication in the CMF group. Nearly all patients in the CMF arm of the parent Protocol B-23 used medication to control nausea throughout each 14-day course of therapy, whereas patients in the AC arm used such therapy for only 3 days after each course but reported nausea (grade 3) 2-4% more often (Fisher, 2001). The use of medication may explain why patients treated with CMF did not report nausea on the QOL questionnaire significantly more often than AC patients. The same results were seen for vomiting. Alopecia was also reported in the parent study at a much higher rate among AC patients (about 80% versus about 40% among CMF patients). Patient self-reported rates were higher than institution-reported rates in both arms. For example, at the beginning of cycle 2, 95% of AC patients and 71% of CMF patients were bothered by hair loss. At cycles 3 and 4 the rate among AC patients had decreased to 85% and then 77%, whereas the rate among CMF patients was 78-79% at the start of cycles 3 and 4. Therefore, the rates based on self-report were more comparable than it would appear based on the institutional reporting, and large differences between the groups are short-term.

The self-reported rates of diarrhea were elevated among placebo as compared to tamoxifen patients, a result that is inconsistent with institution-reported rates (which showed about 1% increase in tamoxifen patients). Unpublished data from the NSABP Breast Cancer Prevention Trial (BCPT) indicate that tamoxifen-treated patients may indeed experience a decrease in diarrhea as compared to placebo patients, according to both self-report and institution-report. The effect sizes in the BCPT were not as large as seen in B-23 self-report, however, so the magnitude of effect reported here (odds ratio 0.32) should be viewed with caution.

While no previous studies were found that compared quality of life in CMF- and AC-treated patients, several previous studies have evaluated the quality of life of breast cancer patients receiving adjuvant CMF. However, these studies either addressed long-term effects of CMF treatment (Joly, 2000) or were performed in node-positive patients. The B-23 study addresses quality of life effects during and immediately after chemotherapy in node-negative patients.

The studies of CMF in node-positive patients included two studies conducted by the IBCSG (Hurny, 1996). The first study was for pre-menopausal and peri-menopausal patients (n=1158 for quality of life evaluation), and the second was for post-menopausal patients (n=940 for quality of life evaluation). Both studies compared different regimens of CMF; the second study included a tamoxifen-only treatment arm. Quality of life was evaluated with individual items for physical well-being, mood, appetite, perceived adjustment/coping, as well as a checklist for emotional well-being, the Befindlichkeits-Skala. Assessments were performed before treatment, 2 months after the start of treatment, and every 3 months for 2 years (and after recurrence). They found that all treatment groups in both studies revealed a substantial improvement in quality of life when nearing the completion of treatment, and an even larger improvement after treatment.

Interestingly, patients appeared to suffer a reduced quality of life due to the anticipation of future chemotherapy (compared to similarly-treated patients who were not scheduled for further treatment). In particular, the QOL scores at 3 months of patients assigned to 3 months of CMF were better than those of patients assigned to 6 months of CMF. This might suggest that patients in the B-23 study who were assigned AC would experience improved QOL near the end of treatment as compared with patients on CMF, for whom treatment is scheduled to last an additional 12 weeks. This is not directly testable in Protocol B-23 because there was no assessment performed immediately after the completion of chemotherapy.

The "ZEBRA" study, a collaboration between Zeneca Pharmaceuticals, the German Breast Cancer Group and the University of Freiberg, also studied the quality of life of node-positive breast cancer patients (Jonat, 1998). The study compared CMF to goserelin in 1466 pre/peri-menopausal node positive stage II patients. The results of the ZEBRA study have not been reported at the time of this manuscript preparation.

In this report we conclude that, with the exception of a few specific symptoms and the time course of the SF-36 vitality scale, quality of life did not differ between chemotherapy treatment arms. The conclusions are complicated by several factors. First, the timing of chemotherapy cycles and quality of life administration was different between the two regimens. In addition, the treatment schedules were different for patients who underwent lumpectomies: in the CMF arm, chemotherapy was delayed during radiotherapy, while in the AC arm, radiotherapy was performed after chemotherapy. Many comparisons are possible, but one study can address only a few. Nonetheless, it is our belief that the comparisons described in this report are sufficiently comprehensive, especially in light of the resounding non-significance of most of the statistical tests performed, to justify the conclusion that no substantial differences in quality of life were apparent in the B-23 protocol. This conclusion might also appear to have been compromised by a loss of statistical power due to the early stopping of the trial and the high rate of non-compliance. However, the sample size achieved (57 in the AC arm and 45 in the CMF arm included in the primary AUC analysis) was adequate to rule out a moderate difference in the primary outcome (effect size 0.56) with 80% power.

The Canadian Medical Association Steering Committee on Clinical Practice Guidelines for the Care and Treatment of Breast Cancer wrote[5] that both AC and CMF are acceptable regimens, and the choice between them should partly depend on quality of life considerations. Protocol B-23 indicates that the choice may be based on practical considerations and patient or physician preferences, without concern for severe disadvantages in terms of quality of life with either regimen.

ACKNOWLEDGMENT

We thank members of the NSABP Behavioral and Health Outcomes Committee for useful discussions of these results. More will be added here...

REFERENCES

[5] (refc 1998)

Brady MJ, Cella DF, Mo F, Bonomi AE, Tulsky DS, Lloyd SR, Deasy S, Cobleigh M, Shiomoto G. Reliability and validity of the Functional Assessment of Cancer Therapy-Breast quality-of-life instrument. J Clin Oncol 1997 Mar;15(3):974-86

Cella, DF, Tulsky, DS, Gray, G et al., Functional assessment of cancer therapy scale: development and validation of the general measure. *J Clinical Oncology*, 11:570-579, 1993.

Little, R. and Rubin, D. (1987) Statistical Analysis with Missing Data, Wiley, New York.

Fisher, B. (2001)

FIGURE LEGENDS

Figure XX. The proportion of completed questionnaires (of those expected) at each assessment time point for each chemotherapy treatment group. The line through the center of each boxplot is the median value, and the upper and lower ends are at the 25^{th} and 75^{th} percentiles, respectively. Solid and dashed lines connect the medians for the AC and CMF groups, respectively. The proportions are approximately the same between treatment groups.

Figure WW. Boxplots of the FACT-B scores at each assessment time point for each chemotherapy treatment arm. There is no difference in scores between chemotherapy treatments. For comparison, the mean FACT-B score among breast cancer patients (all stages) in a validation study was 112.8 (Brady, 1997).

Figure SSS. Boxplots of the SF-36 vitality score at each assessment time point for each chemotherapy treatment arm. AC patients experienced a larger drop in vitality during treatment, and a steeper increase after the conclusion of treatment.

Figure ZZ. The odds ratios for the effects of chemotherapy on each symptom of the symptom checklist and 7 symptoms from the FACT-B, shown with 99.8% confidence intervals (adjusted for multiple comparisons). Values greater than 1 indicate an increase in the odds of the symptom for patients on CMF therapy. Only diarrhea and bladder problems were significantly different, with CMF patients experiencing a greater frequency of these symptoms.

Table XX. Chemotherapy and Quality of Life assessment schedule (in the absence of radiation therapy). The number of expected forms is the number of surviving patients who had not reported a breast cancer recurrence or second primary cancer prior to that assessment time point.

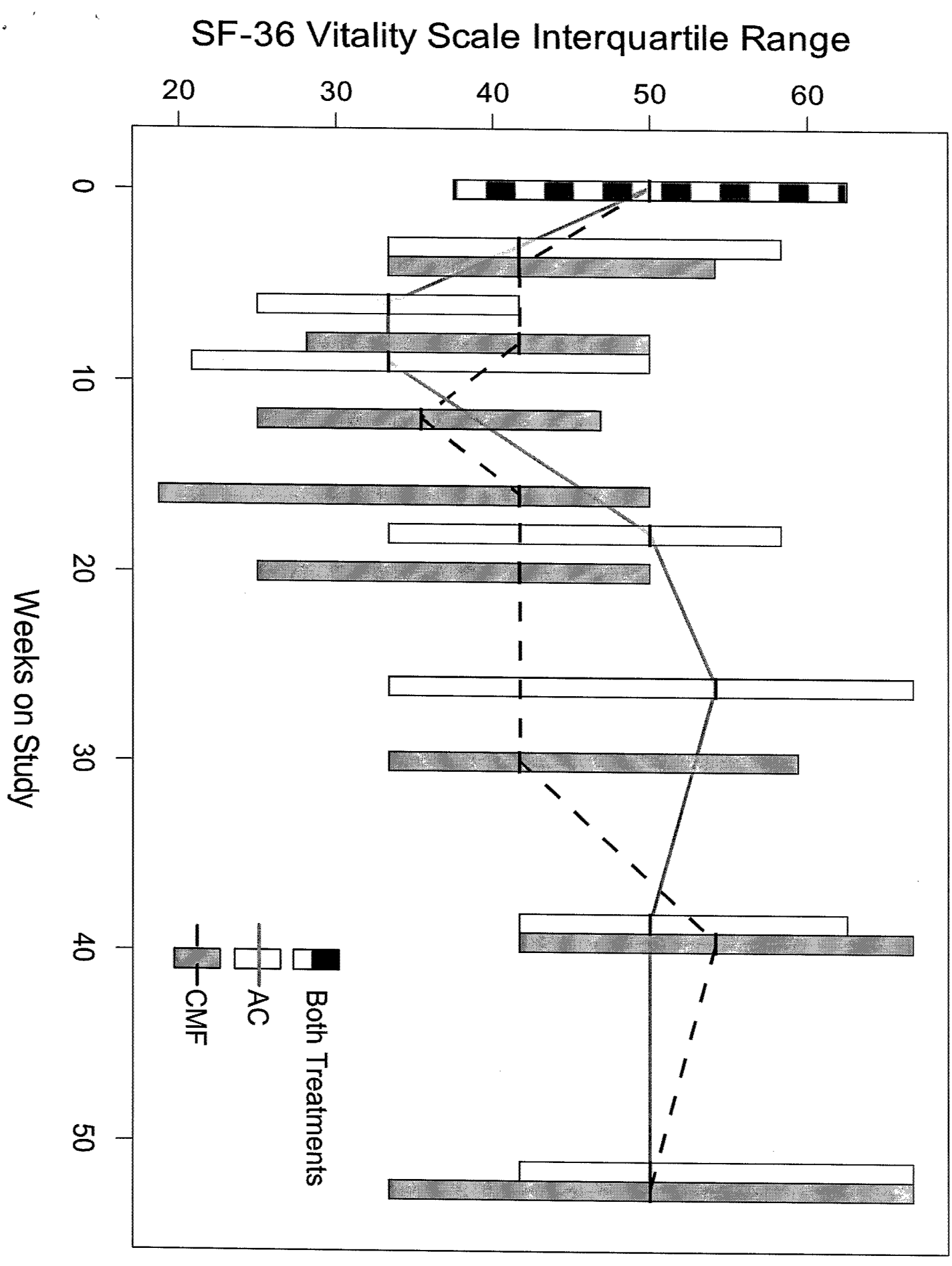
		Week after randomization													
		Baseline	3	4	6	8	9	12	16	18	20	26	30	39	52
AC	Start of Cycle	1	2		3		4								
	Quality of life assessment	√	√		√		√			√		√		√	√
# submitted forms		78	57		65		61			49		47		39	42
# expected forms		87	87		87		85			85		85		84	84
% submitted		90%	66%		75%		72%			58%		55%		46%	50%
CMF	Start of Cycle	1		2		3		4	5		6				
	Quality of life assessment	√		√		√		√	√		√		√*	√	√
# submitted forms		73		65		60		56	55		55		17	40	42
# expected forms		82		82		81		81	80		80		32	77	76
% submitted		89%		79%		74%		69%	69%		69%		53%	52%	55%

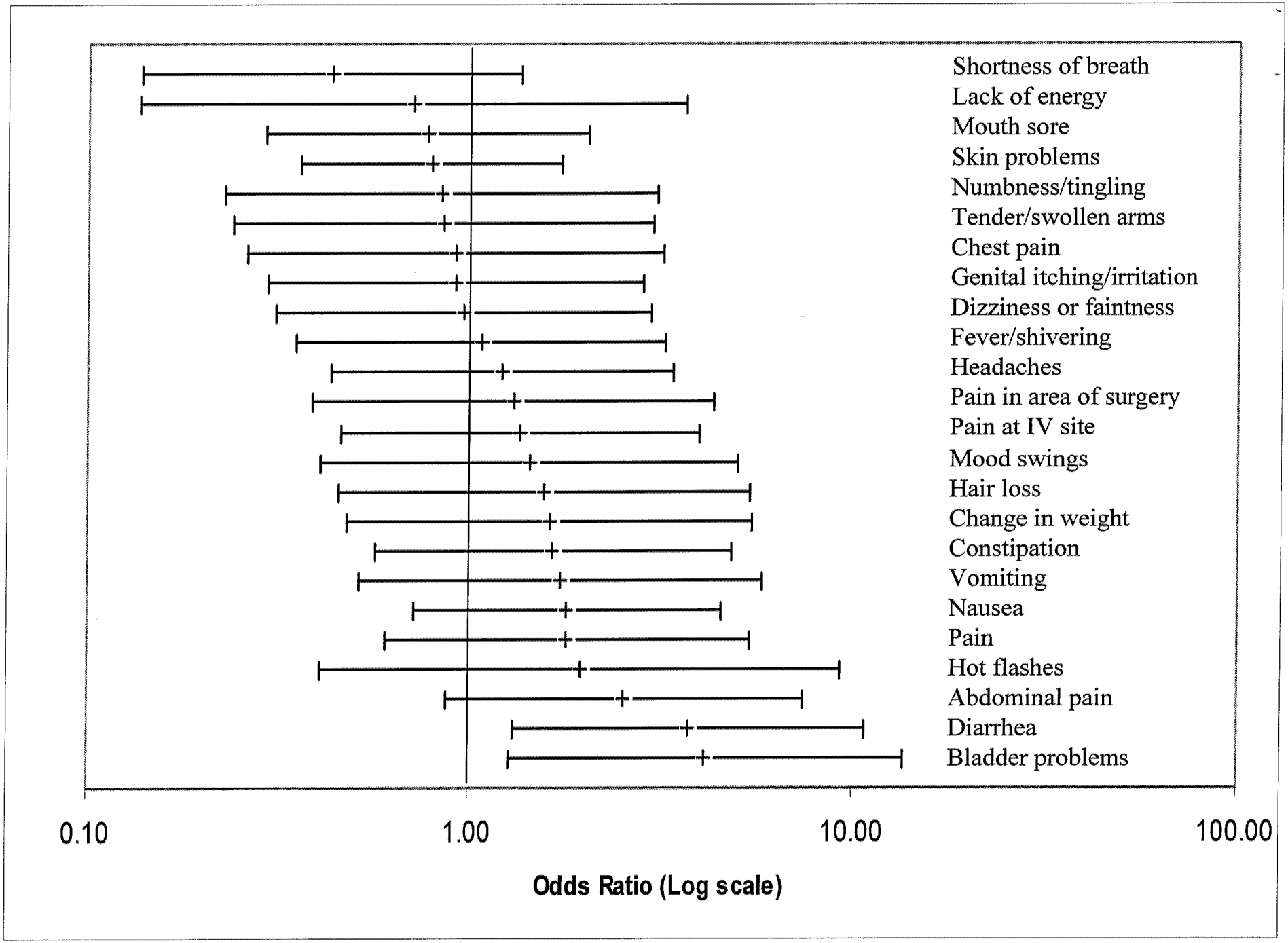
* Mastectomy group only.

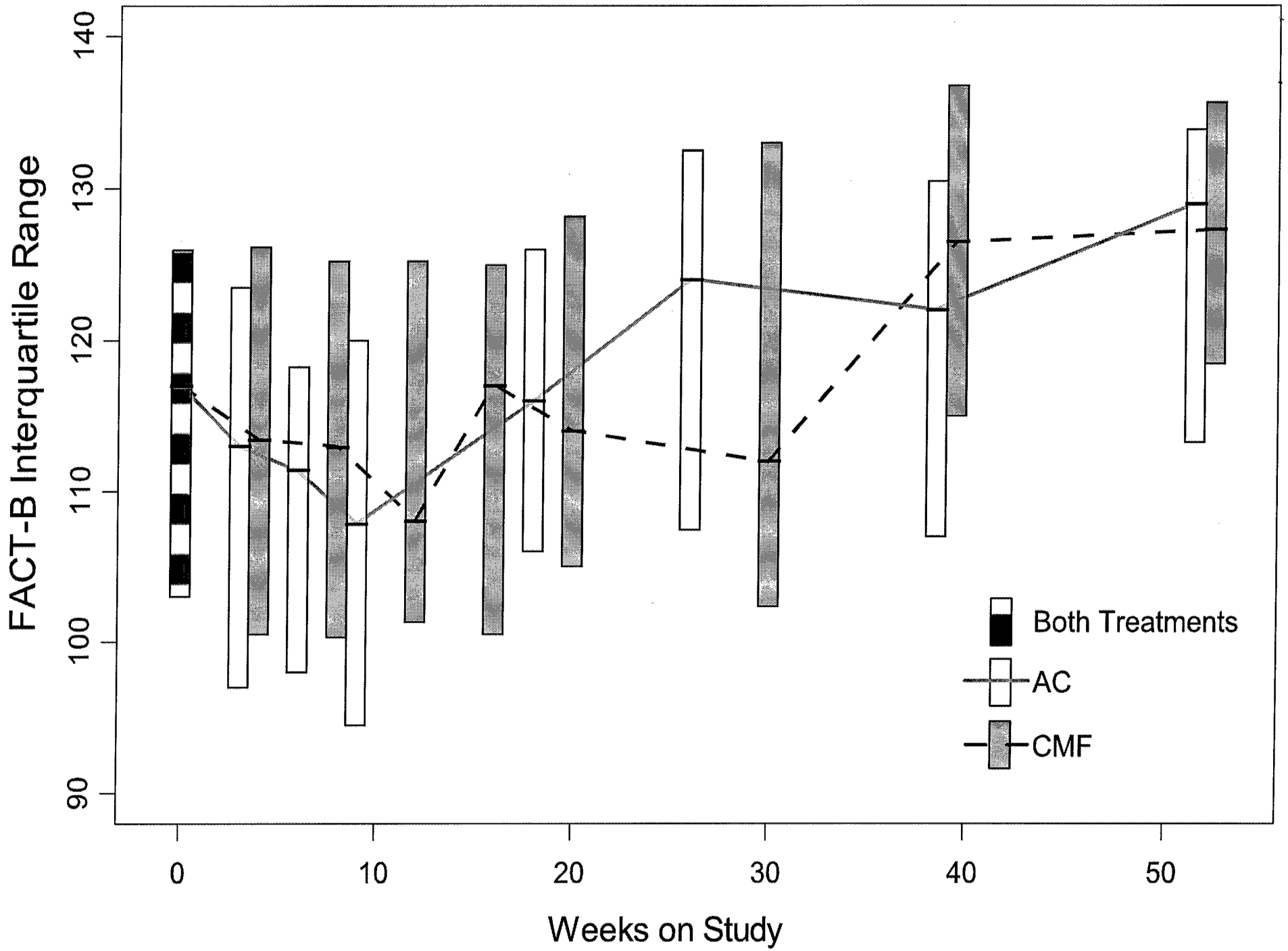
Table YY. For each instrument or item, the first column gives the number of CMF patients who recovered to baseline (row 1) or did not (row 2) at one year after randomization. The second column gives the same numbers for AC patients. The p-values for these comparisons are in the third row. They are based on the Fisher exact test.

	CMF and AC													
	FACT-B		Symptom checklist		SF-36 vitality		HRQOL linear rating scale		Return to normal activities		General health		Comparative health	
Recovered to baseline	29	24	15	19	21	19	29	27	36	37	23	21	34	28
Did not recover	10	15	24	20	17	20	9	11	3	2	16	17	2	1
p-value	.3		.5		.7		.8		1		.8		1	

[6] refc
[7] refc


SF-36 Vitality Scale Interquartile Range
20
30
40
50
60
0
10
20
30
40
50
Weeks on Study
CMF
AC
Both Treatments


Shortness of breath
Lack of energy
Mouth sore
Skin problems
Numbness/tingling
Tender/swollen arms
Chest pain
Genital itching/irritation
Dizziness or faintness
Fever/shivering
Headaches
Pain in area of surgery
Pain at IV site
Mood swings
Hair loss
Change in weight
Constipation
Vomiting
Nausea
Pain
Hot flashes
Abdominal pain
Diarrhea
Bladder problems
0.10
1.00
10.00
100.00
Odds Ratio (Log scale)


FACT-B Interquartile Range
140
130
120
110
100
90
0
10
20
30
40
50
Weeks on Study
Both Treatments
AC
CMF

Determining the Feasibility and Usefulness of Microelectronic Adherence Monitoring Compared to Pill Counts and Self-Reports in a Large, Multicenter Chemoprevention Trial

Richard Day, Ph.D., David Cella, Ph.D., Patricia A. Ganz, M.D., Mary B. Daly, M.D., Ph.D. Julia Rowland, Ph.D., Janet Wolter, M.D.

Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA; Center on Outcomes, Research and Education, Evanston Northwestern Healthcare and robert H. lurie Comprehensive Cancer Center of Northwestern University; Evanston, IL; Jonsson Comprehensive Cancer Center and the Schools of Medicine and Public Health, UCLA, Los Angeles, CA; Fox Chase Cancer Center, Philadelphia, PA; Office of Cancer Survivorship, National Cancer Institute, Bethesda, MD.; Rush Presbyterian-St. Luke's Medical Center, Chicago, IL.

running head: Microelectronic Adherence Monitoring

Corresponding author:
Richard Day, Ph.D.
Department of Biostaistics
Graduate School of Public Health
University of Pittsburgh
Pittsburgh, PA 15261
(412) 624-4077(p)/624-9969(f)
rdfac@vms.cis.pitt.ed

Supported by public health service grants from the National Cancer Institute (NCI-U10-CA-37377/69974) and a career development award from the Department of Defense (DAMD17-97-1-7058)

Abstract

The results of an adherence monitoring substudy are presented from the The National Surgical Adjuvant Breast and Bowel Project's (NSABP) Breast Cancer Prevention Trial (BCPT). The BCPT was a large, multicenter chemoprevention trial in which women at high-risk for breast cancer were given a daily dose of 20mg of tamoxifen or a placebo. Ninety-seven participants from four collaborating centers were followed for six months using three separate methods of adherence monitoring: self-reports, pill counts, and pill caps containing a microelectronic monitoring device. We found acceptable levels of compliance to the daily medication schedule in 90-94% of the study participants and high levels of agreement across all three methods of monitoring medication adherence. We conclude by reviewing certain key aspects of research design and treatment agents that make microelectronic monitoring more or less useful and cost effective in clinical trials.

key words: adherence, compliance, prevention, breast cancer, electronic monitoring

Introduction

The National Surgical Adjuvant Breast and Bowel Project's (NSABP) P-1 Study, the Breast Cancer Prevention Trial (BCPT), was designed to test the efficacy of the antiestrogen drug tamoxifen in preventing breast cancer, fractures and coronary heart disease in healthy women at high-risk for breast cancer. This study was conducted with funds primarily from the National Cancer Institute (NCI), assisted by support from the National Heart, Lung and Blood Institute and the National Institute of Arthritis and Musculoskeletal and Skin Diseases. The P-1 study evolved from a series of studies demonstrating the efficacy of tamoxifen in the prevention of breast cancer systemic recurrence (1-4) and the reduction of contralateral breast cancers in women with early stage breast cancer (1,3,5,6). Other studies demonstrated the benefits of tamoxifen in the lowering of serum cholesterol and increasing bone mineral density in postmenopausal women with breast cancer (7-12) and in the reduction in cardiac events in women with early-stage breast cancer (13).

The P-1 study was a randomized, placebo controlled trial. It was carried out at 119 nucleus clinical centers in the United States and Canada. P-1 recruitment was completed in September 1997 and consisted of 13,388 high-risk women, aged 35-80 years-old, who were randomized to tamoxifen (20 mg per day) or placebo and were scheduled to continue their assigned treatment for 5 years. All participants were to be evaluated at 3 and 6 months during their first year in the study, at six month intervals during the remaining 4 years, and then annually through their seventh year on study.

The findings of the P-1 study were disclosed early (Spring 1998) and participants were notified of their treatment status. Initial findings (14) showed a 49% reduction in the occurrence of invasive breast cancer and a 50% reduction in noninvasive breast cancers among high-risk women. Tamoxifen did not alter the average annual rate of ischemic heart disease; however, a reduction in hip, radius (Colles') and spine fractures was observed. The rate of endometrial cancer (RR=2.53) and rates of stroke, pulmonary embolism, and deep-vein thrombosis were elevated in the tamoxifen group. Women taking tamoxifen also reported an increased frequency of vasomotor and gynecological symptoms and problems of sexual functioning (15,16).

At an early point in the P-1 study, concern was expressed about potential nonadherence to therapy as a threat to the integrity of the trial and the ability to establish the true chemopreventive efficacy of tamoxifen for reduction in the rates of invasive breast cancer, myocardial infarction and bone fractures in high-risk women. Since this was the first study testing tamoxifen in healthy, high-risk women, there was a lack of information about rates of tamoxifen adherence in a preventive framework. It was known that adherence to medical treatment represents a significant problem in a variety of diseases and could result in reductions in statistical power serious enough to affect the evaluation of trial data (17). A clinical report by Waterhouse (18) suggested relatively high levels of adherence to tamoxifen therapy within a treatment context. Similarly, unpublished pill count and self-report data from the NSABP B-14 study indicated a relatively low (10-15%) rate of nonadherence in a group of women treated with tamoxifen for primary breast cancer. However, it was argued that the motivation for adherence in treatment versus chemopreventive settings was likely to be quite different. In the chemopreventive setting people are asked to change their routine behavior for intangible and uncertain future benefits. Concern was expressed that, within a chemopreventive context, even mild side-effects, including the perceived side-effects of placebos, may be sufficient to trigger nonadherence in the well person (15).

Active and placebo treatments used in the P-1 study were distributed in bottles containing a 3-month supply of two hundred 10 mg tablets. Women were provided with two bottles of tablets to cover the period between 6-month follow-up visits. Pill counts and staff assessments based on participant self-reports were built into the P-1 study to estimate treatment adherence. It was known, however, that these techniques of monitoring adherence often tend to underestimate the problems in adherence when they occur, and are usually unable to provide detailed data about the pattern of nonadherence (19). A number of studies in the literature suggested that electronic monitoring of presumptive dosing, which "time stamps" each opening of the pill bottle, provides a more reliable measure of presumed adherence (20-24). However, there was concern about the cost and feasibility of electronic monitoring in a large, multicenter trial. To address these concerns, we designed a four institution substudy to compare three methods of adherence monitoring in the P-1 study: self-report, pill count, and electronic monitoring. The objectives of the substudy were threefold: to test the feasibility and cost-effectiveness of using an electronic monitoring device to measure medication adherence in P-1 study participants; to estimate medication adherence rates for a series of P-1 participants using electronic monitoring data; and, to compare the estimated adherence rates derived from electronic monitoring data to the data obtained from pill counts and participant self-reports.

Study Design and Materials

Subjects - A cohort of 97 women, from four collaborating P-1 institutions (Rush-Presbyterian, UCLA, Fox Chase, Georgetown), participated in the substudy (Table 1). Consecutive P-1 study participants were selected for the study without knowledge of their treatment status (tamoxifen or placebo). IRB committees in each of the collaborating centers approved the substudy. Participants were aware that they were taking part in an adherence monitoring project and signed a separate informed consent. The women were followed-up for adherence using all three methods of monitoring at 3 and 6 months.

Procedures - The Aprex Corporation provided the medication event monitors, software and cap reader used in the research (18,24). Monitor reading was carried out centrally at Rush Presbyterian St. Lukes Hospital in Chicago. A small number of monitors were held in reserve by the centers. This permitted monitors returned on follow-up to be express mailed to Chicago for computerized reading. Replacement monitors were immediately returned to the collaborating centers by express mail. Data from monitor readings were sent to the University of Pittsburgh where they were cleaned and transferred to a database. Data on pill counts and participant self-reports were obtained from the NSABP Data Center and included in the same study database.

Substudy data required extensive summarization prior to analysis. Pill counts were collected on the Treatment Follow-Up Form (TFUF). This required center staff to determine the expected number of pills taken based on start and end dates of the medication period and then to determine whether the number of tablets in the bottle was less than, greater than or equal to the expected number of tablets. Patient self-reports were collected on the Adherence Follow-Up Form (AFUF). This form asked the center staff to interview participants regarding their treatment adherence over the 4 preceding weeks at each follow-up examination, then report on items such as percentage of tablets taken, overall pattern of adherence, and the primary adherence problems of the participant. Different center staff usually completed the

TFUF and AFUF. The output from the electronic monitor provided the dates of the period of observation, a count of the total number of cap openings and daily count of cap openings, the specific time that the cap was opened and closed and the elapsed time since the cap was last opened.

For the purpose of comparing the three adherence monitoring methods, each dosing cycle was set to the longest comprehensive period recorded by any one of the techniques. Prior to the initiation of the trial, clinical estimates derived from what was known about the pharmacokinetics of tamoxifen suggested >75% adherence (i.e., taking tamoxifen an average of 3 of 4 days) would probably be sufficient to maintain adequate medication levels once a therapeutic blood level had been achieved. Therefore, the data from each information source were converted for primary analysis into a binary adherence scale that rated the participant's adherence at each follow-up as "sufficient" (100+% to 76% of drug taken) or "insufficient" (75% to 0%). Only the AFUF formulated its adherence estimates in precisely this manner. The TFUF simply provided an estimate of the number of pills missed over a dosing cycle. For this analysis, it was assumed that pill misses occurred in a random fashion across the dosing cycle. This assumption was generally confirmed by studying a small series of non-adherent participants using data from the monitor output. The monitor data provide a record of cap openings, but cannot assure that the medication was actually ingested. Study participants were asked to take two (10mg) pills daily. That meant that a single opening could reflect either a full (20mg) or a half (10mg) dose. This type of participant-initiated dose reduction was reported in the P-1 study, usually by participants who suspected that the medication was causing uncomfortable side-effects. Similarly, two cap openings could imply an inadvertent overdose or it could be the result from the participant taking half the prescribed medication at two different times during the day.

Statistical Methods - Comparisons of continuous variables were carried out using one and two-way ANOVAs or Kruskal-Wallis and Friedman tests depending upon whether data distributions were approximately normal or not. Tests of proportions were carried out using a chi-square statistic; an exact test was substituted when expected cell values were very small. Data reduction carried out on questionnaire items made it possible to use a kappa statistic (25) to calculate final reliabilities between the monitoring systems.

Results

A total of 14,506 participant days were assessed with electronic monitors, 7962 days in participant months 1-3 and 6544 days in participant months 4-6. At the 3 and 6 month follow-up examinations, 7% of the days monitored had 0 cap openings, 91% had 1 cap opening, and 2% had two cap openings. Only 10 of the 14,506 days monitored over the 6 months of the study had three or more cap openings. Table 2 presents a summary of the participant days of data collected by center. There were no statistically significant differences in the mean or median days monitored between the collaborating centers for either time period.

Forty-seven of the participants were assigned to the tamoxifen arm of the trial and 50 to placebo. There were no statistically significant differences in the proportion of participants assigned to each trial arm across the collaborating centers (Table 3). Of the three monitoring techniques, the pill count data were the most complete, followed by the self-report and the electronic monitoring data (Table 4). Missing data points for the electronic monitoring

technique were clustered at the six month follow-up and occurred with a significantly greater frequency in two of the collaborating centers (Table 2 and 4). With regard to the 18 missing data points, at least 6 were due to an electronic malfunction, resulting in an estimated monitor failure rate of 3.1%. Another 3 missing data points were due to participants who insisted on transferring the tablets out of the bottle and the remaining 9 were the result of unexplained circumstances. Eleven (61%) of the 18 women with missing electronic monitoring data points were assigned to the tamoxifen arm of the trial; however, this association is not of sufficient magnitude to reach statistical significance (X^2,1df = 1.418, p=0.234).

In the P-1 study, participants remained eligible for follow-up whether or not they were taking their assigned treatment. In order to be considered "off trial" and, therefore, lost to follow-up, it was necessary for a participant to formally withdraw their consent to take part in the study. None of the participants in the adherence study were off study through the 6 month follow-up examination, although, at least 7 participants, 4 in the tamoxifen and 3 in the placebo arm, were known to be off treatment.

Table 5 shows that all three monitoring techniques agreed that overall adherence rates were high for the women participating in the P-1 study. There were no statistically significant differences between the centers with regard to the overall adherence rates provided by different monitoring techniques.

Table 6 displays the absolute percentages and kappa statistics for the agreement between the three monitoring techniques at the 3 and 6 month follow-up points. No statistically significant differences were found between individual centers with regard to absolute percentage agreement or kappa statistics at the three or six month follow-up examinations. Treatment status (tamoxifen or placebo) was not associated with levels of agreement between the three monitoring agreements.

Discussion

This is the first report of adherence to a chemopreventive agent in a large scale, North American multicenter cancer prevention trial. Our essential motivation was to obtain practical, comparative experience with electronic monitoring devices and to determine the extent to which other techniques were comparable and adequate in specified settings. Despite the weaknesses of the present study, such as the small cohort size and our inability to continue monitoring beyond the earliest dose cycles, certain definite conclusions emerge from this work.

First, concern was expressed that the introduction of a side-effect producing drug like tamoxifen among a cohort of otherwise healthy, high-risk women might carry significant risk of nonadherence. In fact, during the first six months of treatment, overall compliance in both treatment arms appeared to be quite good, ranging from 90-94%, depending upon the method of monitoring used, in spite of the fact that tamoxifen-related side-effects (e.g., vasomotor and gynecological symptoms) had already surfaced (16). These data suggest, on a short-term basis at least, that acceptable adherence to a daily pill dosing regimen can be expected from motivated individuals at risk for a serious disease. It should be noted that the participants in this substudy were early trial entrants, who were both younger and at higher risk than the final P-1 study cohort. However, the proportion of participants known to be off treatment (.0722,

95%CI: 029-.143) at the time of the six-month follow-up does not differ from expected levels for the overall P-1 study cohort (15,16). Concern has also been expressed that the participants' knowledge that they were taking part in an adherence substudy may have increased their overall levels of medication compliance. However, all of the women taking part in the P-1 study were regularly questioned about their medication-taking behavior and were asked to return unused study tablets to their local clinical staff. To this extent, all of the women participating in the P-1 study were equally aware that they were being regularly monitored for medication adherence.

Second, high levels of overall agreement were observed between all three methods of adherence monitoring. In other words, the electronic monitors were not contributing information that could not be obtained from other, more traditional techniques. In terms of expense, the electronic monitors, like pill counts and self-reports, required the commitment of professional staff resources for the collection, reading and processing of the data. In addition, estimates provided by company representatives in March, 1993 indicated that a larger adherence substudy using electronic monitors in 10 collaborating centers with a total of 1000 participants would cost a minimum of $160,000 per year for monitors, cap readers and software. Implementation of electronic monitoring at that time for the anticipated P-1 cohort of 16,000 women would have cost a minimum of $2.1 million per year or approximately $14 million over the projected life of the trial. At the same time, it should be noted that the cost of electronic monitoring has been substantially reduced over the last 18 months (personal communication, Dr. John Urquhart, Chief Scientist, AARDEX Ltd/APREX Corp).

Third, our experience also suggests that there are certain important aspects of research design and the treatment agent that make electronic monitoring systems more or less useful and cost effective for clinical trials. We recommend that any investigator who is considering the use of a electronic monitoring system carefully review the following aspects of the proposed trial:

a. **The pharmacokinetic characteristics of the medication being tested.** One of the virtues of tamoxifen as a preventative agent is that therapeutic levels, once established, can be maintained over time under a relatively flexible dosing routine. This may be contrasted with medications having a relatively short half-life and requiring a rigid dose schedule in order to maintain therapeutic levels in the blood. From a statistical point of view, the issue with regard to the two types of medications is the likelihood of losing sufficient power to reject the null hypothesis. Clearly, there is a greater likelihood of losing statistical power when a trial involves the latter as compared to the former type of medication. Hence, the more rigid the required dosing schedule for the experimental agent, the greater the importance of adherence monitoring and the value of implementing multiple monitoring techniques, perhaps, including microelectronic devices.

b. **The physical characteristics of the agent being tested and pill distribution system.** In the P-1 study, participants were asked to take two 10 mg tablets daily. As a consequence, one reported cap opening could represent either a 100% or 50% dose and two cap openings might represent a 100% dose or an unknown level of over-medication. This meant that none of the three monitoring techniques used in the study could be considered a "gold standard" for the others. In addition, a small but not

insignificant number of our participants developed their own techniques for insuring adherence that involved removing the pills from the P-1 study bottle and placing them in other containers (e.g., daily medication boxes). Finally, the tamoxifen pills used in the P-1 study were distributed after the six month follow-up in two 200 tablet bottles lasting three months each. From a cost perspective, this meant that each participant required two electronic monitors between follow-up examinations. The only alternative would be to ask the participants to transfer the electronic monitor from the first to the second pill bottle.

c. **The fundamental objectives of the clinical trial.** A clinical trial may be narrowly focused on issues of biological efficacy or it may be concerned with issues of biological efficacy within the practical context of treatment delivery and client adherence. In the former type of trial active interventions to support the participants' adherence make good methodological sense and electronic monitoring systems can play an important role. The electronic monitoring method, for example, produces an abundance of information that can be directly shared with the trial participant in order to reinforce acceptable adherence or to develop strategies designed to overcome less than adequate adherence. In contrast, large scale chemoprevention trials like the P-1 study take a more passive attitude towards adherence monitoring and are interested in testing whether a particular agent is effective within the practical, real world context of the dosing behavior exhibited by high-risk, but otherwise healthy, women living in the general population. This attitude recognizes that the relatively passive adherence monitoring experienced by trial participants is likely to be far more active than the routine levels of monitoring that will be exercised if the agent is approved for general use. In this context, the electronic monitoring system tends to provide an overload of information which is often ignored or grossly simplified.

In summary, our experience suggests that electronic monitoring systems are not necessarily an optimal technique for adherence monitoring in large-scale chemoprevention trials like the P-1 study. Instead, electronic monitoring systems appear best suited for more intensive, small scale clinical trials that are focused primarily on issues of biological efficacy and are able to implement active forms of adherence monitoring and participant support.

Acknowledgements

The authors wish to acknowledge the assistance of the AARDEX Ltd./APREX Corp., Zug CH & Union City, CA, USA; John Urquhart, MD, FRCP (Edin), Chief Scientist, AARDEX Ltd./APREX Corp; Joyce A. Cramer, BS, VA Medical Center, New Haven, CT; Pamela Witcher, Ph.D., Georgetown University; and Joyce Ho, Ph.D., University of Pittsburgh.

References

1. Fisher B, Costantino J, Redmond C, Poisson R, Bowman D, Couture J et al. A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor-positive tumors. N Engl Med 1989; 320:479-84.

2. Controlled trial of tamoxifen as single adjuvant agent in management of early breast cancer. Analysis at six years by Nolvadex Adjuvant Trial Organization. Lancet 1985; 1:836-40.

3. Adjuvant tamoxifen in the management of operable breast cancer: the Scottish trial. Report from the Breast Cancer Trial Committee. Scottish Cancer trials Office (MRC), Edinburgh. Lancet 1987; 2:171-5.

4. Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy. 133 randomized trials involving 31,000 recurrences and 24,000 deaths among 75,000 women. Early Breast Cancer Trialists' Collaborative Group. Lancet 1992; 339:1-15,71-85.

5. Cyclophosphamide and tamoxifen as adjuvant therapies in the management of breast cancer. CRC Adjuvant Breast Trial Working party. Br Cancer 1988; 57:604-7.

6. Fornander T, Rutqvist LE, Cedermark B, Glas U, Mattsson A, Silfversward C et al. Adjuvant tamoxifen in early breast cancer: occurrence of new primary cancers. Lancet 1989; 1:117-20.

7. Rutqvist LE, Mattsson A. Cardiac and thromboembolic morbidity among postmenopausal women with early breast cancer in a randomized trial of adjuvant tamoxifen. The Stockholm Breast Cancer Study Group. J Natl Cancer Inst 1993; 85:1398-406.

8. Bertelli G, Pronzato P, Amoroso D, Cusimano MP, Conte PF, Montagn G et al. Adjuvant tamoxifen primary breast cancer: influence on plasma lipids and antithrombin II levels. Breast Cancer Res Treat 1988; 12:307-10.

9. Rossner S, Wallgren A. Serum lipoproteins after breast cancer surgery and effects of tamoxifen. Atherosclerosis 1984; 52:339-46.

10. Bruning PF, Bonfrer JM, Hart AA, de Jong-Bakker M, Linders D, var Loon J, et al. Tamoxifen, serum lipoproteins and cardiovascular risk. Br J Cancer 1988; 58:497-9.

11. Love RR, Newcomb PA, Wiebe DA, Surawicz TS, Jordan VC, Carbone PP, et al. Effects of tamoxifen therapy on lipid and lipoprotein levels in postmenopausal patients with node-negative breast cancer. J Natl Cancer Inst 1990; 82:1327-32.

12. Love RR, Mazess R, Barden H, Epstine S, Newcomb P, Jordan V, et al. Effect of tamoxifen on bone mineral density in postmenopausal women with breast cancer. N Engl J Med 1992;326:852-856.

13. Rutqvist LE, Mattsson A. Cardiac and thromboembolic morbidity among postmenopausal women with early-stage breast cancer in a randomized trial of adjuvant tamoxifen. The Stockholm Breast Cancer Study Group. J Natl Cancer Inst 1993; 85:1398-406.

14. Fisher B, Costantino JP, Wickerham DL, Redmond CK, Kavanah M, Cronin WM et al. Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. J Natl Cancer Inst 1998, 90:1371-1388.

15. Ganz PA, Day R, Ware Jr. JE, Redmond C, Fisher B. Base-line quality-of-life assessment in the National Surgical Adjuvant Breast and Bowel Project Breast Cancer Prevention Trial. J Natl Cancer Inst 1995; 87:1372-82.

16. Day R, Ganz PA, Costantino JP, Cronin WM, Wickerham DL, Fisher B. Health-related quality of life and tamoxifen in breast cancer prevention: a report from the National Surgical Adjuvant Breast and Bowlel Project P-1 Study. J Clin Oncol 1999, 17:2659-2669.

17. Cramer JA, Spliker B (Eds): Patient Compliance in Medical Practice and Clinical Trials. New York: Raven Press, 1991.

18. Waterhouse DM, Calzone KA, Mele C, Brenner DE. Adherence to oral tamoxifen: a comparison of patient self-report, pill counts, and microelectronic monitoring. J Clin Oncology 1993; 11:1189-97.

19. Spilker B. Methods of assessing and improving patient compliance in clinical trials. In JA Cramer & B Spilker (Eds): Patient Compliance in Medical Practice and Clinical Trials. New York: Raven Press, 1991, 37-56.

20. Peterman AH and Cella DF. Adherence issues among cancer patients. In: Shumaker SA, Schron EB, Ockene JK, Mc Bee W (eds.), The Handbook of Health Behavior Change, 462-511. New York: Springer Publishing Company, 1998.

21. Cramer JA. Microelectronic systems for monitoring and enhancing patient compliance with medication regimens. Drugs 49: 321-7, 1995.

22. Urquhart J. The electronic medication event monitor – lessons for pharmacotherapy. Clin Pharmacokinet 32: 345-356, 1997.

23. Kastrissios H, Blaschke TF. Medication compliance as a feature in drug development. Ann Rev Pharmacol Toxicol 37: 451-75, 1997.

24. Lee R, Nicholson PW, Souhami RL, Deshmukh AA. Patient compliance with oral chemotherapy as assessed by a novel electronic technique. J Clin Oncol, 1992, 19:1007-1013.

25. Fleiss JL. The Design and Analysis of Clinical Experiments. New York: John Wiley & Sons, 1986.

Table 1
Descriptive Characteristics by Collaborating Center of the BCPT Participants Recruited for the Adherence Substudy

Variable	Collaborating Center				Totals
	1	2	3	4	
Cohort Size	25	22	24	26	97
Age					
35-49 yrs	9 (36%)	10 (45%)	13 (54%)	11 (42%)	43 (44%)
50-59 yrs	7 (28%)	9 (41%)	9 (38%)	14 (54%)	39 (40%)
60+ yrs	9 (36%)	3 (14%)	2 (8%)	1 (4%)	15 (16%)
Ethnicity					
White	25 (100%)	22 (100%)	23 (96%)	25 (96%)	95 (98%)
Black	0	0	1 (4%)	0	1 (1%)
Other	0	0	0	1 (4%)	1 (1%)
Relative Risk					
$\leq$ 2.0	2 (7%)	2 (9%)	0	0	4 (5%)
2.01-3.00	4 (15%)	0	2 (8%)	4 (15%)	10 (10%)
3.01-5.00	7 (30%)	11 (50%)	7 (29%)	14 (54%)	39 (40%)
5.01-10.00	8 (33%)	5 (23%)	8 (34%)	6 (23%)	27 (28%)
$\geq$ 10.01	4 (15%)	4 (18%)	7 (29%)	2 (8%)	17 (17%)

Table 2

Mean and Median Participant Days Electronically Monitored by Collaborating Center and Time Period

Time Period and Variable	1	2	3	4	Totals
Months 1-3					
Complete data	23	22	23	25	93
Missing data	2	0	1	1	4
Mean days	88.4	87.1	89.9	84.9	87.5
SD	16.4	18.9	2.5	19.3	15.7
Median days	92	92	90	89	90
Months 3-6					
Complete data[1]	20	15	23	25	83
Missing data[1]	5	7	1	1	14
Mean days	73.1	87.8	81.8	84.1	81.8
SD	36.3	13.5	25.8	25.6	26.2
Median days	90	91	90	90	90

1. Exact p for 2x4 table = 0.015

Table 3

Number and Percent of Adherence Study Participants Assigned to Tamoxifen and Placebo by Collaborating Center

Variable	1	2	3	4	Totals
Tamoxifen	12 (48%)	10 (45%)	12 (50%)	13 (50%)	47 (48%)
Placebo	13 (52%)	12 (55%)	12 (50%)	13 (50%)	50 (52%)
Totals	25 (100%)	22 (100%)	24(100%)	26(100%)	97(100%)

Table 4
Proportion of Data Completed Using the Pill Counts,
Self-Report and Electronic Monitors by Collaborating Center and Time Period

Time Period and Variable	1	2	3	4	Totals
Months 1-3					
Pill Count	100% (25/25)	100% (22/22)	100% (24/24)	100% (26/26)	100% (97/97)
Self-Report	100% (25/25)	100% (22/22)	96% (23/24)	92% (24/26)	97% (94/97)
Electronic Monitor	92% (23/25)	100% (22/22)	96% (23/24)	96% (25/26)	96% (93/97)
Months 3-6					
Pill Count	100% (25/25)	100% (22/22)	100% (24/24)	100% (26/26)	100% (97/97)
Self-Report	100% (25/25)	100% (22/22)	100% (24/24)	100% (26/26)	100% (97/97)
Electronic Monitor	80% (20/25)	68% (15/22)	96% (23/24)	96% (25/26)	86% (83/97)

Table 5
Proportion of Study Participants Estimated to Show Sufficient Adherence (>75% of Tablets) by Different Monitoring Techniques, Time Periods and Centers

Time Period and Variable	1	2	3	4	Totals
Months 1-3					
Pill Count	100% (25/25)	90% (20/22)	96% (23/24)	92% (24/26)	95% (92/97)
Self-Report	100% (25/25)	95% (21/22)	96% (22/23)	92% (22/24)	96% (90/94)
Electronic Monitor	96% (22/23)	95% (21/22)	96% (22/23)	92% (23/25)	95% (88/93)
Months 3-6					
Pill Count	88% (22/25)	100% (22/22)	96% (23/24)	92% (24/26)	94% (91/97)
Self-Report	84% (21/25)	95% (21/22)	88% (21/24)	92.3% (24/26)	90% (87/97)
Electronic Monitor	80% (16/20)	100% (15/15)	91% (21/23)	92% (23/25)	90% (75/83)

Table 6

Absolute Proportion and Unweighted Kappa Statistics for Overall Agreement Between Different Adherence Monitoring Techniques by Time Period

Measure of Agreement	Months 1-3		Months 4-6	
	Electronic Monitor	Pill Count	Electronic Monitor	Pill Count
Absolute Proportion of Agreement				
Pill Count	.957	n/a	.975	n/a
Self-Report	.967	.989	.963	.959
Unweighted Kappa				
Pill Count	.577	n/a	.820	n/a
Self-Report	.650	.883	.781	.729

Scaling symptoms relevant when using hormonal therapies to prevent breast cancer: Results from the first prevention study of the National Surgical Adjuvant Breast and Bowel Project (NSABP)

Chih-Hung Chang[1]

David Cella[1]

Patricia A. Ganz[2]

Richard Day[3]

(and others per NSABP)

[1] Center on Outcomes, Research and Education (CORE), Evanston Northwestern Healthcare, and Institute for Health Services Research and Policy Studies and Robert H. Lurie Comprehensive Cancer Center, Northwestern University

[2] Schools of Medicine and Public Health, and the Johsson Comprehensive Cancer Center, University of California at Los Angeles

[3] Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Address Correspondence to:

Chih-Hung Chang, Ph.D.
Center on Outcomes, Research and Education
Evanston Northwestern Healthcare
1001 University Place, Suite 100
Evanston, IL 60201
Phone: (847) 570-7311
Fax: (847) 570-8033
Email: chchang@northwestern.edu

Abstract

Purpose: To conduct scaling and psychometric evaluation on symptom checklist (SCL) data collected at the baseline (pre-treatment) and 12-month assessments of the first NSABP Breast Cancer Prevention Trial (P-1).

Patients and Methods: Data came from responses of 11,064 women recruited into a study of 20 mg daily tamoxifen versus placebo for the prevention of breast cancer in high risk women. Exploratory factor analyses were first conducted on a random sample of 4000 women to establish initial factor structures using 12-month data. Baseline data were then used for confirmatory factor analyses. The remaining sample was further divided randomly into two data sets. Data on each set were then grouped by age (35-49, 50-59, and $\geq$ 60 years) and treatment (tamoxifen and placebo) for both exploratory and confirmatory factor analyses.

Results: Eight clusters of symptoms were identified and confirmed.

Conclusion:

Psychometric Assessment of the Symptom "Checklist" from the NSABP Breast Cancer Prevention Trial (P-1)

This article reports the results of scaling and psychometric evaluation on symptom checklist (SCL) data from 11,064 women collected at the baseline (pre-treatment), 3, 6, 12, 24 and 36-month assessments of the National Surgical Adjuvant Breast and Bowel Project (NSABP) Breast Cancer Prevention Trial (P-1).[1] The SCL data are gathered on page 3 of the quality of life questionnaire for that study. The form has come to be called a "checklist," because it originated from the Postmenopausal Estrogen/Progesterone Intervention (PEPI) trial conducted by the National Heart Lung and Blood Institute (NHLBI). The NHLBI version of the questionnaire was indeed a checklist of symptoms, where participants endorsed only their presence or absence. The NSABP modification was an expanded list of symptoms and a two-part response format in which women first indicate presence or absence of the symptom, then rate the severity ("bother") on a five-point scale, where "0"=not at all and "4"=extremely. This SCL questionnaire has never been subjected to systematic psychometric analysis. As a result, it remains unclear exactly how to use it in interpreting trial outcome. Ganz et al.[2] were forced to use a simple summing of endorsed symptoms, diminishing the contribution this scale can make to outcome reporting on the trial. A systematic approach to scaling will create clinically meaningful and psychometrically sound subscales that will improve the interpretability and ultimate value of the SCL data collected on the trial.

PATIENTS AND METHODS

Participants

This report covers the baseline health-related quality of life (HRQL) examination and the first 36 months of follow-up data on 11,064 women recruited over 24 moths (June 1, 1992, to May 31,

1994). Demographic, medical and treatment factors (baseline age in years, hysterectomy, both ovaries removed, menstrual periods stopped, treatment group),

medical history of angina, heart attack, heart failure, heart murmur, high blood pressure, TIA, stroke, vascular problems, arthritis, bone fractures, Osteoporosis, gallstones or gallbladder disease, diabetes, liver disease, thyroid trouble, TB, and prior malignancy information were collected.

Symptom Checklist

The NSABP investigators initially reviewed and considered the PEPI trial symptoms provided by Dr. Sally Shumaker for use in the prevention study. Beginning with a list of symptoms from the PEPI trial checklist, additional relevant symptoms were added for comprehensive coverage. Items were reviewed and approved by a 10-member expert quality of life panel (P.G., and D.C., member). A total of 42 symptoms were retained in the final questionnaire. The items assess common physical and psychological symptoms, with a particular emphasis upon symptoms associated with menopause (e.g. hot flashes, vaginal dryness) and tamoxifen use (e.g., vaginal discharge). The response format selected was drawn from the experience of the University of Rochester Community Clinical Oncology Program (G. Morrow, personal communication).

On the symptom checklist (SCL), women are asked if they have encountered any of the 42 listed symptoms. If yes, they further endorse severity ("how much the problem bothered you") for each endorsed symptom on a 5-point Likert-type scale (0=Not at all; 1=Slightly; 2=Moderately; 3=Quite a bit; 4=Extremely). The present and severity responses to each symptom were combined to create a new 6-point scale for subsequent analyses: 0=No; 1=Yes, Not at all; 2=Yes, Slightly; 3=Yes, Moderately; 4=Yes, Quite a bit; 5=Yes, Extremely). In addition, each patient also answers the Medical Outcomes Study Short Form-36 (SF-36),[3] and the Center for Epidemiologic Studies - Depression Scale (CES-D).[4]

Item Reduction/Scale Construction Analyses

The underlying factor structure of the 42-item SCL data (yes/no and severity) from baseline (N=11,064) and 12-month follow-up data were investigated via both exploratory and confirmatory factor analyses for traditional factor analysis for scale construction;

Item response data and scale scores were analyzed using the sample as a whole and separately by the three different age groups (35-49, 50-59, and ≥ 60 years) and two treatment arms (placebo vs. tamoxifen). Figure 1 details the steps taken to derive interpretable symptom clusters.

Insert Figure 1 about here

Briefly,.....

The stability of obtained scale(s)/underlying construct(s) across three different age groups and two treatment arms were evaluated via confirmatory factor analyses. After scale construction is completed, concurrent validity was examined [redacted], using the profile and summary scores of the SF-36 to define different groups. Health status summary scores on the SF-36 will be used after scale construction being completed to test concurrent validity;

Newly-constructed SCL scale scores will be analyzed for their responsiveness to treatment arm (tamoxifen versus placebo) in an intent-to-treat analysis.

Exploratory Factor Analysis

In the initial scale development stage, a factor analysis technique was employed to explore the underlying structure of these 42 symptoms. Because the baseline responses were recorded before administration of any drug, we used the 12-month data for these exploratory analyses. This will

capitalize on whatever symptom "clustering" occurs as result of taking the drug; yet still allows us to return to the baseline data to check scale performance. Exploratory factor analysis[5] enabled us to identify the redundancy in a set of correlated variables and to reduce the set to a smaller number of derived variables called factors. It is a way of item grouping that allows us to investigate the underlying structure of a correlation matrix. In exploratory factor analysis, such as was the case in this study, there are a series of steps to take and decisions to make. Among them are: which extracting method to use; how many factors to retain; which rotational technique to use; and what criteria to set for identifying items that mark a factor. These issues are discussed briefly below.

Principal components analysis (PCA),[7,8] is one of the most commonly used procedures. In PCA, linear combinations of the observed variables are formed. The components are estimated to represent the variances of the observed variables. For example, the first principal component accounts for the largest amount of variance, and the second component explains the next largest amount of variance and is uncorrelated to the first one. Each component has an eigenvalue, which is the amount of variance accounted for by the component. To decide the "proper" number of factors for exploratory factor analysis, one common rule is to retain for rotation any eigenvalue (or latent root) greater than 1.0, or Kaiser-Guttman criterion. The number of factors to retain can be further evaluated using Cattell's scree test[9] which plots the incremental variance accounted for by each successive factor to determine the point at which the explained variance levels out. We will use this scree technique to explore multiple optimal factor solutions, as described below.

Factor rotation, orthogonal or oblique, is usually required to find a best (simple structure) solution, which makes the retained factors more interpretable and meaningful. Orthogonal rotation using the varimax procedure,[10] in which factors are kept uncorrelated, to produce reasonable simple structure is most commonly used in exploratory factor analysis. To determine the clusters of items, the items with

the highest factor loadings are selected for the scale. Factor loadings are generally considered meaningful when they exceed .30 or .40. Items with small (below 0.30) factor loadings will be omitted from the obtained factors.

In summary, to detect patterns in the correlation matrix among the 42 symptoms, 12-month responses of the women to the checklist will be analyzed using PCA with orthogonal rotation to a Varimax criterion. The number of factors and item groupings will be reviewed and determined by the scree test and factor loadings. These criteria advance how many factors could be constructed and to which factor the item belongs. Scales will be then constructed for each distinctive factor, with items selected using the criteria as described above. The SPSS statistical package will be used to facilitate these steps. Because the sample size is so large (N=11,064), we can conduct multiple factor analyses on independent samples. We propose to conduct the exploratory factor analyses on the sample as a whole and three age groups of patients (35-49; 50-59; >= 60 years).[2] Factor structures obtained from separate analysis will be compared. While we do not expect perfect conformity of item composition between any of the exploratory comparisons, the comparison with the most consistently matching items will be selected for further refinement.

Confirmatory Factor Analysis

[To add]

Reconciliation of item composition of scales from two approaches

The primary focus of this study was to determine which items to keep in each factor analysis-derived scale, after applying both exploratory and confirmatory factor analyses. It should be noted that the items dropped from the two approaches were studied carefully later. They could either be re-added into a scale for conceptual or practical reasons (usually with

minimal loss to reliability and precision), or they can be set aside and analyzed separately for clinical interest.

In summary, the development of the scales began with exploratory factor analyses and followed by confirmatory factor analyses and then to reconcile the subscale composition. There was likely to be a good deal of similarities from different analyses and any remaining inconsistencies can be evaluated and rectified on conceptual (i.e., clinical) grounds. The reduction of the SCL data into meaningful and usable components, which was important for its utmost utility as a clinical and research tool, can optimally be achieved. [redacted]

[redacted]

[redacted]

Determining Validity

All validity analyses were conducted after the initial scaling procedures described in this proposal being carried out and the specific nature of the subscales being defined. Concurrent validation will be possible by virtue of the concurrent collection of health status data (Medical Outcomes Study Short Form-36),[3] and depression symptom data (Center for Epidemiologic Studies - Depression Scale).[4] Dr. Richard Day has agreed to assume responsibility for these analyses.

Finally, and of most clinical/public health interest, the longitudinal data will be analyzed for tamoxifen versus placebo differences in an intent-to-treat analysis using appropriate multivariate models and handling of missing data. [redacted]

RESULTS

The results from all the analyses were summarized in Table 1. Eight interpretable cluster of symptoms were identified with each has 2 or 3 items. [Do we want to create another table to show these factors with items]

Confirmatory factor analyses on these identified factors showed high goodness-of-fit indexes. [Say more here and provide statistics]

CONCLUSION

Eight reproducible scales across age groups, treatment arms, and time.

21 of 42 items included.

Additional 21 items require expert input for disposition.

REFERENCES

Day R, Ganz PA, Cosantino JP, et al: Health-related quality of life and Tomaxifen in breast cancer prevention: A report from the National National Surgical Adjuvant Breast and Bowel Project P-1 Study. J Clin Oncol 17:2659-2669, 1999.

Ganz PA, Day R, Ware JE, Redmond C, et al: Base-line quality-of-life assessment in the National Surgical Adjuvant Breast and Bowel Project breast cancer prevention trial. J Natl Cancer Inst 87(18):1372-1382, 1995.

Ware JN, Kosinski M, Keller SD: SF-36 physical and mental health summary scales: A user's manual. Boston, MA: The Health Institute, New England Medical Center, 1994.

Ranloff LS: The CES-D scale: A self-report depression scale for research in the general population. Appl Psycho Meas 1: 385-401.

Harman HH: Modern factor analysis (3rd ed.). Chicago, University of Chicago Press, 1976.

Spearman C: "General intelligence" objectively determined and measured. American Journal of Psychology 15: 201-293.

Pearson K: On lines and planes of closest fit to systems of points in space. Phil. Mag., ser. 2, 6: 559-572, 1901.

Hotelling H: Analysis of a complex of statistical variables into principal components. JEP 24: 417-441, 498-520.

Cattell RB: The scree test for the number of factors. Multivariate Behavioral Research 1: 245-276.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. Psychometrika, 23, 187-200.

Rasch G: Probabilistic models for some intelligence and attainment test. Chicago, University of

Chicago Press, 1980.

Wright BD, Masters GN: Rating scale analysis. Chicago, MESA Press, 1982.

Wright BD, Stone MH: Best test design. Chicago, MESA Press, 1979.

Rasch G: On general laws and the meaning of measurement in psychology, Vol. 4. (pp. 321-334). Proceedings of 4th Berkeley Symposium on Mathematical Statistics. Berkeley, CA: University of California Press, 1961.

Rasch G: An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 19: 49-57, 1966.

Rasch G: An individualistic approach to item analysis. In P. Lazarsfeld and N.V. Henry (Eds.), Reading in Mathematical Sciences (pp. 89-107). Chicago: Science Research Association, 1966.

Wright B D: Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service, 1968.

Wright B, Panchapakesan N: A procedure for sample-free item analysis. Educational and Psychological Measurement 29: 23-48, 1969.

Masters GN: A Rasch model for partial credit model. Psychometrika 47: 149-174,

Andrich D. Application of a psychometric rating model to ordered categories which are scored with successive integers. Applied Psychological Measurement, 1978a, 2: 581-594.

Andrich D: A rating formulation for ordered response categories. Psychometrika 43: 561-573, 1978b.

Gehlert, S., & Chang, C.-H. (1998). Factor structure and dimensionality of the Multidimensional Health Locus of Control Scale in measuring patients with epilepsy. Journal of Outcome

Measurement, 2(3), 179-190.

Chang, C.-H. (1998). Confirming test structure and measurement characteristics. Rasch measurement Transactions, 12(1), 622-623.

Prieto, L., Alonso, J., Lamarca, R., & Wright, B. D. (1998). Rasch measurement for reducing the items of the Nottingham Health Profile. Journal of Outcome Measurement, 1998, 2, 285-301.

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? Journal of Outcomes Measurement, 2(3), 266-283.

Linacre, J. M. (1999). Learning from principal components analysis of residuals (http://www.rasch.org/comet.htm).

Linacre, J. M. & Wright, B. D. (2000). WINSTEPS Rasch model computer program. Chicago: MESA Press.

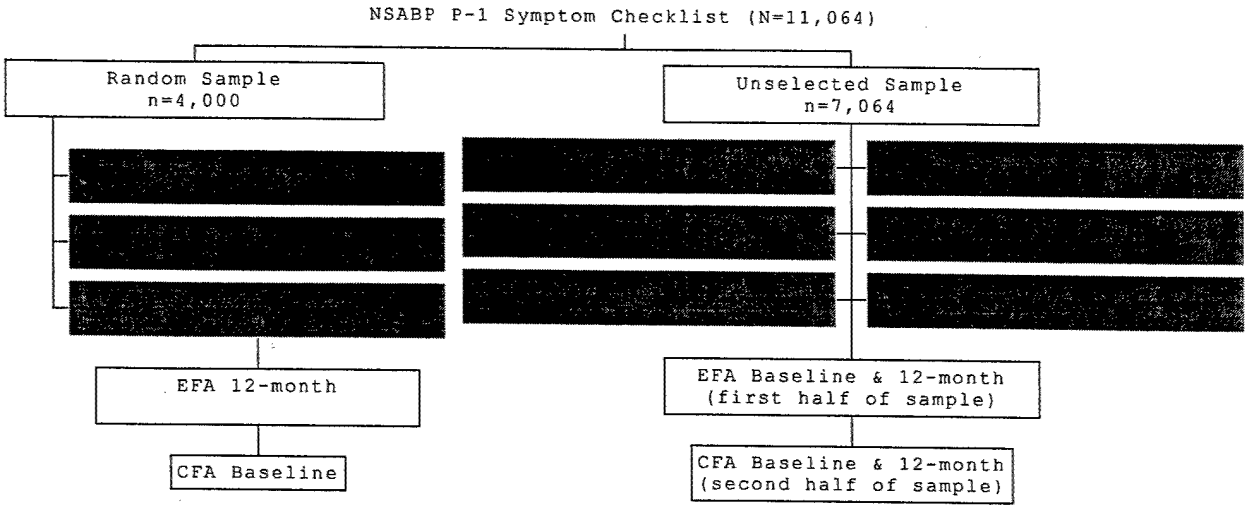


Figure 1. Sample Selection and Analysis Flow

Table 1. Exploratory Factor Analysis Results

Symptom	12-month		12 -Month						Baseline						% membership
			35-49		50-59		60+		35-49		50-59		60+		
	T	P	T	P	T	P	T	P	T	P	T	P	T	P	
38. Difficulty concentrating	●	●	●	●	●	●	●	●	●	●	●	●	●	●	100
39. Easily distracted	●	●	●	●	●	●	●	●	●	●	●	●	●	●	100
30. Forgetfulness	●	●	●	●	●	●	●	●	●	●	●	●	●	●	100
33. Tendency to take nap; stay in bed	●	●				●		●							29
37. Avoidance of social affairs	●	●				●					●	●			36
36. Tendency toward accident															0
19. Joint pain	●	●	●	●	●	●	●	●	●	●	●	●	●	●	100
22. Muscle stiffness	●	●	●	●	●	●	●	●	●	●	●	●	●	●	100
18. General aches and pains	●	●	●	●	●	●	●	●	●	●	●	●	●	●	100
21. Swelling of hands or feet						●						●			14
41. Numbness, tingling													●		7
34. Night sweats	●	●	●	●	●	●	●		●	●	●	●	●		86
1. Hot flashes	●	●	●	●	●	●	●		●	●	●	●	●		86
35. Cold sweats	●		●	●	●	●			●	●			●		57

Symptom	12-month		12 -Month						Baseline						% membership
			35-49		50-59		60+		35-49		50-59		60+		
	T	P	T	P	T	P	T	P	T	P	T	P	T	P	
42. Early awakening															0
5. Vomiting	●	●	●	●	●	●	●	●	●	●	●	●	●	●	100
4. Nausea	●	●	●	●	●	●	●	●	●		●	●	●	●	93
6. Diarrhea	●		●	●	●	●		●	●	●		●		●	71
23. Difficulty breathing	●	●	●					●	●		●		●	●	57
29. Feeling of suffocation	●	●	●					●			●		●	●	50
20. Chest pains	●	●	●					●	●						36
13. Vaginal dryness	●	●	●	●	●	●	●	●	●	●	●	●	●	●	100
14. Pain with intercourse	●	●	●	●	●	●	●	●	●	●	●	●	●	●	100
8. Difficulty with bladder control (when laughing or crying)	●	●	●	●	●	●	●	●	●	●	●	●	●	●	100
9. Difficulty with bladder control (at other times)	●	●	●	●	●	●	●	●	●	●	●	●	●	●	100
24. Dry mouth															0
17. Ringing in ears		●		●		●		●	●	●					43
3. Blind spots, fuzzy vision			●	●						●				●	29
7. Constipation	●										●	●			21

Symptom	12-month		12 -Month						Baseline						% membership
			35-49		50-59		60+		35-49		50-59		60+		
	T	P	T	P	T	P	T	P	T	P	T	P	T	P	
40. Dizziness, faintness		●				●		●	●	●	●	●		●	57
2. Headaches	●		●			●						●			29
15. Cramps	●	●		●					●						29
16. Breast sensitivity/ tenderness	●	●		●					●						29
32. Short temper													●	●	14
31. Excitability													●	●	14
26. Weight loss	●		●		●	●	●								36
25. Weight gain	●	●	●	●			●	●	●	●	●	●	●	●	86
28. Decreased appetite					●		●								14
27. Unhappy with the appearance of my body	●	●	●	●		●		●	●	●	●	●	●	●	86
10. Vaginal discharge	●	●	●	●	●	●	●	●	●		●				71
12. Genital itching/ irritation	●		●	●	●	●	●		●			●			57
11. Vaginal bleeding or spotting	●	●		●		●		●			●	●			50

* Check in box indicates relevant factor loading ≥ .40.